CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation

Chujie Zheng[†], **Yong Liu**[‡], **Wei Chen**[‡],
Yongcai Leng[‡], **Minlie Huang**[†*]
[†]The CoAI group, DCST, Institute for Artificial Intelligence,
[†]State Key Lab of Intelligent Technology and Systems,
[†]Beijing National Research Center for Information Science and Technology,
[†]Tsinghua University, Beijing 100084, China
[‡]Sogou Inc., Beijing, China
chujiezhengchn@gmail.com,

Abstract

The capacity of empathy is crucial to the success of open-domain dialog systems. Due to its nature of multi-dimensionality, there are various factors that relate to empathy expression, such as communication mechanism, dialog act and emotion. However, existing methods for empathetic response generation usually either consider only one empathy factor or ignore the hierarchical relationships between different factors, leading to a weak ability of empathy modeling. In this paper, we propose a multi-factor hierarchical framework, CoMAE, for empathetic response generation, which models the above three key factors of empathy expression in a hierarchical way. We show experimentally that our CoMAE-based model can generate more empathetic responses than previous methods. We also highlight the importance of hierarchical modeling of different factors through both the empirical analysis on a real-life corpus and the extensive experiments. Our codes and used data are available at `https://github.com/chujiezheng/CoMAE`.

1 Introduction

Empathy, which refers to the capacity to understand or feel what another person is experiencing (Rothschild, 2006; Read, 2019), is a critical capability to open-domain dialog systems (Zhou et al., 2018b). As shown in previous research, empathetic conversational models can improve user satisfaction and receive more positive feedback in numerous domains (Klein, 1998; Liu and Picard, 2005;

Brave et al., 2005; Fitzpatrick et al., 2017; Liu et al., 2021). Recently, there have also been numerous works devoted to improving the dialog models' ability to understand the feelings of interlocutors (Rashkin et al., 2019; Lin et al., 2019; Majumder

*Corresponding author.

aihuang@tsinghua.edu.cn

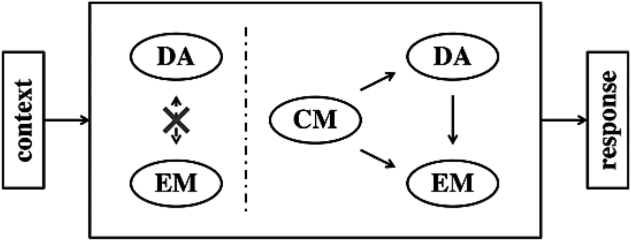


Figure 1: Our proposed hierarchical framework: CoMAE (right). The directed arrows denote dependencies. We also present the framework (left) of EmpTransfo (Zandie and Mahoor, 2020) for comparison.

et al., 2020), which makes the dialog models more empathetic to a certain extent.

However, empathy is a multi-dimensional construct (Davis et al., 1980) rather than merely recognizing the interlocutor's emotion (Lin et al., 2019) or emotional responding (Zhou et al., 2018a). It consists of two broad aspects related to *cognition* and *affection* (Omdahl, 2014; Paiva et al., 2017). The cognitive aspect requires understanding and interpreting the situation of the interlocutor (Elliott et al., 2018), which is reflected in the **dialog act** taken in the conversation (De Vignemont and Singer, 2006), such as questioning (e.g., *What's wrong with it?*), consoling (e.g., *You'll get through this*), etc. The affective aspect relates to properly expressing **emotion** in reaction to the experiences and feelings shared by the interlocutor, such as admiration (e.g., *Congratulations!*), sadness (e.g., *I am sorry to hear that*), etc. Very recently, Sharma et al. (2020) further characterizes the text-based expressed empathy based on the above two aspects as three **communication mechanisms**, which is a more higher-level and abstract factor that relates to empathy expression.

In this paper, we propose a novel framework named **CoMAE** for empathetic response generation (Section 3), which contains the aforementioned three key factors of empathy expression: **Co**mmunication **M**echanism (CM), dialog **A**ct (DA) and **E**motion (EM). Specifically, when model these empathy factors simultaneously, we adopt a **hierarchical** way instead of following previous works that treat multiple factors independently, such like EmpTransfo (Zandie and Mahoor, 2020) that considers both DA and EM (see Figure 1 for comparison). Such approaches hold the hypothesis that different factors are independent of each other, which is intuitively unreasonable. In fact, our empirical analysis (Section 4) on a Reddit corpus (Zhong et al., 2020) shows that there are obvious hierarchical relationships between different factors, which confirms the soundness and necessity of hierarchical modeling.

We then devise a CoMAE-based model on top of the pre-trained language model GPT-2 (Radford et al., 2019) (Section 5), and compare the model performance with different combinations of empathy factors and hierarchical modeling. Automatic evaluation (Section 6.3) shows that combining all the three factors hierarchically can achieve the best model performance. Manual evaluation (Section 6.4) demonstrates that our model can generate more empathetic responses than previous methods. Extensive experiments (Section 6.5) further highlight the importance of hierarchical modeling in terms of the selection and realization of empathy factors.

The contributions of this paper can be summarized in three folds:

- Based on the nature of multi-dimensionality of empathy expression, we propose a novel framework, CoMAE, for empathetic response generation. It hierarchically models three key factors of empathy expression: communication mechanism, dialog act and emotion.

- On top of GPT-2, we devise a CoMAE-based model. Experimental results show that our model can generate more empathetic responses than previous methods.

- We empirically analyze the necessity of hierarchical modeling, and highlight its importance especially in terms of the selection and realization of different empathy factors.

2 Related Work

2.1 Factors Related to Empathy Expression

Empathy is a complex multi-dimensional construct (Davis et al., 1980) which consists of two broad aspects related to *cognition* and *affection* (Omdahl, 2014; Paiva et al., 2017). As shown in Section 1, the two aspects are reflected in the dialog act (DA) taken and the emotion (EM) expressed in the conversation respectively.

Based on the theoretical definition of empathy, Sharma et al. (2020) characterize the text-based expressed empathy as 3 communication mechanisms (CM): emotional reaction (ER) (e.g., *I feel really sad for you*), interpretation (IP) (e.g., *This must be terrifying*, *I also have similar situations*), and exploration (EX) (e.g., *Are you still feeling alone now?*).[1] These communication mechanisms are also applied in the recently proposed task of empathetic rewriting (Sharma et al., 2021).

Besides, Zhong et al. (2020) propose that persona, which refers to the social face an individual presents to the world (Jung, 2016), has been shown to be highly correlated with personality (Leary and Allen, 2011), which in turn influences empathy expression (Richendoller and Weaver III, 1994; Costa et al., 2014). While Zhong et al. (2020) do not explain the explicit connection between persona and empathy expression, they suggest that different speakers may have different "styles" for expressing empathy.

2.2 Empathetic Response Generation

In the past years, empathetic response generation has attracted much research interest (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Zandie and Mahoor, 2020; Sun et al., 2021). Rashkin et al. (2019) suggest that dialog models can generate more empathetic responses by recognizing the interlocutor's emotion. Lin et al. (2019) propose to design a dedicated decoder to respond each emotion of the interlocutor, which makes the generation process more interpretable. Majumder et al. (2020) adopt the idea of emotional mimicry (Hess and Fischer, 2014) to make the generated responses more empathetic. Inspired by the advances in generative pre-trained language models (Radford et al., 2018, 2019), EmpTransfo (Zandie and Mahoor, 2020) uses GPT (Radford et al., 2018) to generate empathetic responses.

Unlike previous works that only consider the EM factor in empathy modeling, EmpTransfo takes both DA and EM into account. The fundamental difference of EmpTransfo from our work lies in two points: (1) our work further considers communication mechanism in modeling empathy, and (2)

[1] As shown in (Sharma et al., 2020), the three communication mechanisms can be properly combined in one utterance. We refer the readers to their original paper for more details about the three communication mechanisms.

we analyze and explore in depth the importance of hierarchically modeling of these empathy factors.

3 CoMAE Framework and Formulation

Our proposed CoMAE framework is shown in Figure 1. CoMAE uses CM as a high-level factor that provides a coarse-grained guidance for empathy expression, and then takes DA and EM to achieve the fine-grained realization. Formally, given the context x, CoMAE divides the generation of the empathetic response y into four steps: (1) predict **CM** C_y conditioned on the context, (2) predict **DA** A_y conditioned on both the context and CoM, (3) predict **EM** E_y based on all the conditions, and (4) generate the final response y. The whole process is formulated as Equation 1:

$$\begin{aligned}\mathbb{P}(y, C_y, A_y, E_y|x) = \mathbb{P}(y|x, C_y, A_y, E_y)\cdot \quad (1)\\ \mathbb{P}(E_y|x, C_y, A_y)\mathbb{P}(A_y|x, C_y)\mathbb{P}(C_y|x).\end{aligned}$$

Note that EM is conditioned on DA, because we intuitively think the expressed emotion is the effect rather than the cause of taking some dialog act. In the other words, one may not adopt the dialog act just for the purpose of expressing some emotion. Hence, realizing the emotion expression as expected is also important in our task, which is the motivation of that we analyze the realization of different factors in Section 6.5.

It is also worth noting that while CoMAE only contains the three factors, such hierarchical framework can be naturally extended to more factors that relate to empathy expression. For instance, Zhong et al. (2020) suggest that persona plays an important role in empathetic conversations. Due to that persona may contain the information about the speaker's style of adopting DA or expressing EM, when integrating persona into empathetic response generation, being conditioned on DA and EM may lead to better performance.

4 Data Preparation and Analysis

While no empathetic conversation corpora provide annotations of diverse empathy factors, there are abundant publicly available resources that make automatic annotation feasible. In this section, we first introduce our used corpus and the resources and tools used in automatic annotation, then we show our empirical analysis to verify the hierarchical relationships between different empathy factors.

4.1 Corpus

Zhong et al. (2020) propose a large-scale empathetic conversation corpus[2] crawled from Reddit. It has two different domains: Happy and Offmychest. The posts in the Happy domain mainly have positive sentiments, while those in the Offmychest domain are usually negative. We adopted their corpus for study for two major reasons: (1) the corpus is real-life, scalable and naturalistic rather than acted (Rashkin et al., 2019), and (2) the manual annotation in (Zhong et al., 2020) shows that most of the last responses are empathetic (73% and 61% for Happy and Offmychest respectively).

4.2 Annotation Resources

Communication Mechanism (CM)[3] Sharma et al. (2020) provide two corpora annotated with CM: TalkLife (`talklife.co`) and Reddit (`reddit.com`), while only the latter is publicly accessible and we thus used the Reddit part. Note that in their original paper, each mechanism is differentiated as three classes of "no", "weak", or "strong". Due to the unbalanced distribution of three classes, we merged "weak" and "strong" into "yes". Finally, we differentiated each mechanism as two classes: "no" or "yes".

Dialog Act (DA)[4] Welivita and Pu (2020) propose a taxonomy of DA (referred as "intent" in the original paper) for empathetic conversations. They first annotate 15 initial types of DA on the ED corpus (Rashkin et al., 2019), and finally obtain 8 high-frequency types of DA with other types merged as others (**8+others**), which are shown in Figure 2.

Emotion (EM)[5] We considered the taxonomy proposed in (Demszky et al., 2020), which contains 27 emotions and a neutral one, because: (1) it has a wide coverage of emotion categories with clear definitions, and (2) the annotated corpus is large-scale and also crawled from Reddit. However, we noted that the original emotion distribution is unbalanced and the too fine-grained taxonomy may lead to the sparsity of partial emotions. Considering the task

[2] https://github.com/zhongpeixiang/PEC

[3] https://github.com/behavioral-data/Empathy-Mental-Health

[4] https://github.com/anuradha1992/EmpatheticIntents

[5] https://github.com/google-research/google-research/tree/master/goemotions

Classifiers	Corpora	# classes	Acc	F1-macro
CM-ER	Reddit	2	81.2	76.9
CM-IP	Reddit	2	85.7	85.7
CM-EX	Reddit	2	96.4	92.5
DA	ED	9	92.0	87.8
EM	Reddit	10	60.5	60.4

Table 1: Performance of the classifiers. "ED" refers to the corpus of EMPATHETICDIALOGUES (Rashkin et al., 2019).

scenario of empathetic conversation, we adopted the clustering results in (Demszky et al., 2020) and modified the original taxonomy as 9 emotions and a neutral one (**9+neutral**), which are also shown in Figure 2. We show the mapping between our adopted emotions and the original emotions in Appendix A.

4.3 Classifiers

We fine-tuned the RoBERTa[6] (Liu et al., 2019) classifiers for CM, DA and EM, whose performance is summarized in Table 1. They all achieve reasonable performance, ensuring the quality of automatic annotation.

However, we noted that the source domain (Rashkin et al., 2019) of the DA classifier is different from the target domain (Reddit). To verify the quality of DA annotation, we recruited three workers from Amazon Mechanical Turk to judge whether the utterance is consistent with the annotated DA. From the utterances that are not annotated with "others", we randomly sampled 25 utterances for each DA (totally 200) to avoid the impact of unbalanced distribution. Finally, the ratio of being judged as consistent is 0.78 with Fleiss' Kappa $\kappa = 0.621$ (Fleiss, 1971), which indicates substantial agreement ($0.6 < \kappa < 0.8$) and that the automatic annotation of DA is also reliable.

4.4 Data Filtering and Annotation

Following the original data split of (Zhong et al., 2020), we first filtered those conversations where there are more than two speakers (about 15%) to ensure that the last utterance is related to the post. We used the aforementioned classifiers to automatically annotate each utterance with DA and EM, and annotate each final response additionally with CM. We found that the last responses that are not annotated with any CM are more likely to

[6] https://huggingface.co/roberta-base

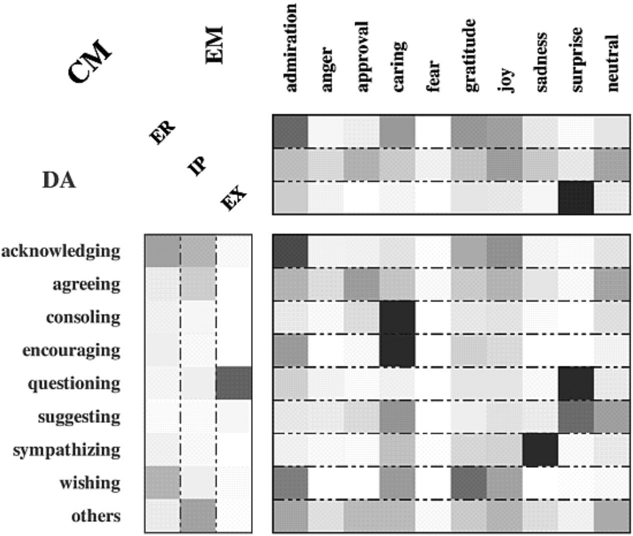


Figure 2: Heat maps of the conditional distributions between the three empathy factors. The **orange / red / blue** maps are the distributions of **DA / EM / EM** conditioned on **CM / CM / DA** respectively.

be non-empathetic, thus we filtered the conversations containing such responses (about 40%). Finally, the sizes of Train / Valid / Test-Happy / Test-Offmychest are 125,963 / 16,371 / 11,136 / 6,413 respectively. We show the detailed statistics of automatic annotation in Appendix B.

4.5 Analysis

In order to verify the hierarchical relationships between the three factors, we counted the distribution frequency of each (X, Y)[7] pair, where (X, Y) is one of the three factor pairs: (CM, DA), (CM, EM), (DA, EM). We approximated the statistical frequency of (X, Y) as their joint probability distribution $\mathbb{P}(X, Y)$. We then normalized $\mathbb{P}(X, Y)$ along the X dimension to obtain the conditional distribution of Y given X: $\mathbb{P}(Y|X)$.

Figure 2 shows the heat maps of the conditional distributions of the three factor pairs. The heat maps reveal obvious patterns of the occurrence of

Y given X. For instance, when one adopts the DA *encouraging*, he usually expresses the EM *caring* instead of *approval* or *joy*. If one expresses empathy with the CM *exploration (EX)*, he almost always adopts the DA *questioning* and expresses the EM *surprise*. Hence, considering the hierarchical relationships between different empathy factors is reasonable and natural, and is also necessary for better empathy modeling.

[7] X or Y is the random variable that represents CM, DA, or EM.

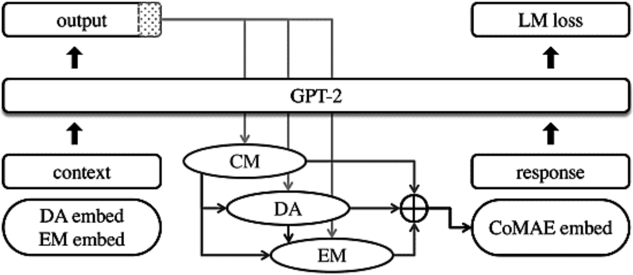


Figure 3: The overall architecture of our CoMAE-

based model. The position and speaker embeddings are omitted for simplicity. The orange dashed block denotes the output hidden state at the last position of the context.

5 Methodology

5.1 Model Architecture

Our devised CoMAE-based model uses GPT-2 as the backbone (Radford et al., 2019). The overall architecture is shown in Figure 3.

Firstly, our model takes the dialog context x as input. The context x is the concatenation of history utterances: $x = (u_1, u_2, \ldots, u_N)$, where N is the length of dialog history. Any two adjacent utterances are also separated by the special token `[EOS]`. Each history utterance u_i contains a sequence of tokens: $u_i = (u_{i,1}, u_{i,2}, \ldots, u_{i,l_i})$, where l_i is the length of u_i. Each utterance u_i is labeled with the corresponding speaker $k_{u_i} \in \{0, 1\}$ (only 2 speakers). We denote the annotated DA and EM of each utterance u_i as $A_{u_i} \in [0, 9)$ and $E_{u_i} \in [0, 10)$ respectively. Suppose that the token id and the position id of $u_{i,j}$ are denoted as $w_{u_{i,j}} \in [0, |\mathcal{V}|)$ ($\mathcal{V}$ is the vocabulary) and $p_{u_{i,j}} \in [0, 1024)$ (the maximum input length is 1024) respectively, the representation of each token $u_{i,j}$ is the summation of the following embeddings:

$$\begin{aligned} \boldsymbol{e}_{u_{i,j}} = \boldsymbol{M}_W \left[w_{u_{i,j}}\right] + \boldsymbol{M}_P \left[p_{u_{i,j}}\right] + \\ \boldsymbol{M}_K \left[k_{u_i}\right] + \boldsymbol{M}_A \left[A_{u_i}\right] + \boldsymbol{M}_E \left[E_{u_i}\right], \end{aligned} \quad (2)$$

where $\boldsymbol{M}_W \in \mathbb{R}^{|\mathcal{V}| \times d}, \boldsymbol{M}_P \in \mathbb{R}^{1024 \times d}, \boldsymbol{M}_K \in \mathbb{R}^{2 \times d}, \boldsymbol{M}_A \in \mathbb{R}^{9 \times d}, \boldsymbol{M}_E \in \mathbb{R}^{10 \times d}$ denote the embedding matrices of word, position, speaker, DA and EM respectively, and $[\cdot]$ denotes the indexing operation. We denote the output hidden states after feeding x into the model as $\boldsymbol{H}_x \in \mathbb{R}^{l_x \times d}$, where l_x is the total length of context x.

Next, we use the hidden state at the last position of the context, $\boldsymbol{h}_x = \boldsymbol{H}_x[-1] \in \mathbb{R}^d$, to hierarchically predict the CM, DA and EM of the target response. We first separately predict[8]

[8] In the mathematical notation used in this paper, we dis-

$\widehat{C}_y^{(i)} \in \{0, 1\}$ for each $i \in \{\mathrm{ER}, \mathrm{IP}, \mathrm{EX}\}$, which indicates whether to adopt the CM i:

$$\boldsymbol{h}_C^{(i)} = \mathbf{F}_C^{(i)}(\boldsymbol{h}_x) \in \mathbb{R}^d, \tag{3}$$

$$\widehat{C}_y^{(i)} \sim \mathbb{P}\left(C_y^{(i)} \middle| x\right) = \mathrm{softmax}\left(\boldsymbol{M}_C^{(i)} \boldsymbol{h}_C^{(i)}\right),$$

$$\widehat{C}_y = \left(\widehat{C}_y^{(\mathrm{ER})}, \widehat{C}_y^{(\mathrm{IP})}, \widehat{C}_y^{(\mathrm{EX})}\right),$$

$$\boldsymbol{e}_{\widehat{C}_y} = \sum_{i \in \{\mathrm{ER}, \mathrm{IP}, \mathrm{EX}\}} \boldsymbol{M}_C^{(i)} \left[\widehat{C}_y^{(i)}\right], \tag{4}$$

where each $\mathbf{F}_C^{(i)}$ is a non-linear layer activated with tanh, and each $\boldsymbol{M}_C^{(i)} \in \mathbb{R}^{2 \times d}$ denotes the embedding matrix of the CM $i \in \{\mathrm{ER}, \mathrm{IP}, \mathrm{EX}\}$. Based on the context x and the predicted CMs $\widehat{C}_y$, we next predict DA:

$$\boldsymbol{h}_A = \mathbf{F}_A\left(\left[\boldsymbol{h}_x; \boldsymbol{e}_{\widehat{C}_y}\right]\right) \in \mathbb{R}^d, \tag{5}$$

$$\widehat{A}_y \sim \mathbb{P}\left(A_y \middle| x, \widehat{C}_y\right) = \mathrm{softmax}\left(\boldsymbol{M}_A \boldsymbol{h}_A\right), \tag{6}$$

where $[\cdot;\cdot]$ denotes vector concatenation and $\mathbf{F}_A$ is a non-linear layer. Note that we share the parameters of DA embeddings with the classification head (Equation 6), which is consistent with the way in GPT-2 (Radford et al., 2019) where the parameters of word embeddings are shared with the LM head (Equation 10). EM is predicted similarly but conditioned additionally on the predicted DA $\widehat{A}_y$:

$$\boldsymbol{h}_E = \mathbf{F}_E\left(\left[\boldsymbol{h}_x; \boldsymbol{e}_{\widehat{C}_y}; \boldsymbol{M}_A\left[\widehat{A}_y\right]\right]\right) \in \mathbb{R}^d, \tag{7}$$

$$\widehat{E}_y \sim \mathbb{P}\left(E_y \middle| x, \widehat{C}_y, \widehat{A}_y\right) = \operatorname{softmax}\left(\boldsymbol{M}_E \boldsymbol{h}_E\right), \tag{8}$$

where $\mathbf{F}_E$ is also a non-linear layer.

Finally, we add all the factors to obtain the fused embedding $\boldsymbol{e}_{\text{CoMAE}}$ that controls the empathy expression of the response:

$$\boldsymbol{e}_{\text{CoMAE}} = \boldsymbol{e}_{\widehat{C}_y} + \boldsymbol{M}_A\left[\widehat{A}_y\right] + \boldsymbol{M}_E\left[\widehat{E}_y\right].$$

The embedding of each input token $\widehat{y}_t$ in the response is as follows:

$$\begin{aligned}\boldsymbol{e}_{\widehat{y}_t} = \boldsymbol{M}_W\left[w_{\widehat{y}_t}\right] + \boldsymbol{M}_P\left[p_{\widehat{y}_t}\right] + \\ \boldsymbol{M}_K\left[k_y\right] + \boldsymbol{e}_{\text{CoMAE}}.\end{aligned} \tag{9}$$

Suppose that the output hidden state corresponding to $\widehat{y}_t$ is $\boldsymbol{s}_t$, then we predict the next token $\widehat{y}_{t+1}$ through the LM head:

$$\begin{aligned}\widehat{y}_{t+1} &\sim \mathbb{P}\left(y_{t+1}\,\middle|\,\widehat{y}_{\leq t}; x, \widehat{C}_y, \widehat{A}_y, \widehat{E}_y\right) \\ &= \operatorname{softmax}\left(\boldsymbol{M}_W \boldsymbol{s}_t\right),\end{aligned} \tag{10}$$

where the parameters of the LM head are shared with the word embedding matrix $\boldsymbol{M}_W$.

5.2 Training

The optimization object contains two parts. One part is the negative log likelihood loss $\mathcal{L}_{\text{NLL}}$ of the target response:

tinguish the ground truth value and the predicted value of a variable X with the symbols X^* and $\widehat{X}$ respectively.

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{l_y}\sum_{t=1}^{l_y} \ln \mathbb{P}\left(y_t^* \middle| y_{<t}^*; x, C_y^*, A_y^*, E_y^*\right),$$

where l_y is the length of the golden response. The other part is the prediction losses of CM $\mathcal{L}_C$, DA $\mathcal{L}_A$, and EM $\mathcal{L}_E$:

$$\mathcal{L}_C = -\sum_{i\in\{\text{ER,IP,EX}\}} \ln \mathbb{P}\left(C_y^{(i)*}\middle| x\right), \quad (11)$$

$$\mathcal{L}_A = -\ln \mathbb{P}\left(A_y^* \middle| x, C_y^*\right), \quad (12)$$

$$\mathcal{L}_E = -\ln \mathbb{P}\left(E_y^* \middle| x, C_y^*, A_y^*\right). \quad (13)$$

The complete optimization object is the summation of the above losses: $\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda\left(\mathcal{L}_C + \mathcal{L}_A + \mathcal{L}_E\right)$, where λ is the weight of the prediction losses. We set λ to 1.0 in our experiments.

5.3 Discussion

It is worth noting that the supervision signals of predictions (from Equation 11 to 13) combined with hierarchical modeling (from Equation 3 to 8) enable the model to establish the connections between the embeddings of the three factors. For instance, in Equation 6, the embedding matrix of DA, $\boldsymbol{M}_A$, is multiplied with $\boldsymbol{h}_A$, which explicitly contains the information of the embedding matrices of CM, $\boldsymbol{M}_C^{(i)}$ (Equation 4 and 5). The case of Equation 8 is similar, where $\boldsymbol{M}_E$ is multiplied with $\boldsymbol{h}_E$ that directly relates to $\boldsymbol{M}_C^{(i)}$ and $\boldsymbol{M}_A$.

Hence, consider two models where one uses hierarchical modeling and the other does not (predicting each factor separately). When the two models are fed with the same empathy factors, saying the triplet (C_y, A_y, E_y) is designated validly, we can expect that the former model has better performance than the latter one. This conjecture will be verified in the automatic evaluation (Section 6.3).

6 Experiments

6.1 Compared Models

We investigated the model performance with different combinations of empathy factors and hierarchical modeling:

(1) **Vanilla**: the GPT-2 model directly fine-tuned on the corpus without adding any empathy factor;

(2) **+CM**, **+DA**, **+EM**: the GPT-2 models equipped with one of the three factors;

(3) **CM || DA**, **CM || EM**, **DA || EM**, **CM || DA || EM**: the models equipped with two or all of the three factors, but predicting each factor separately without hierarchical modeling;

(4) **CM → DA**, **CM → EM**, **DA → EM**, **CM → DA → EM**: the models that are similar to (3) but utilize the hierarchical relationships, where → denotes dependency.

Note that the baseline DA || EM is consistent with EmpTransfo[9] (Zandie and Mahoor, 2020), and CM → DA → EM is exactly our devised model described in Section 5.1.

6.2 Implementation Details

All the models were implemented with PyTorch[10] (Paszke et al., 2019) and the Transformers library[11] (Wolf et al., 2020). We used the pre-trained GPT-

2 with the size of 117M parameters (768 hidden sizes, 12 heads, 12 layers) for all the models. The responses were decoded by Top-p sampling with $p = 0.9$ and the temperature $\tau = 0.7$ (Holtzman et al., 2019). We trained all the models with Adam (Kingma and Ba, 2014) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate was 10^{-4} and was dynamically changed using the linear warmup (Popel and Bojar, 2018) with 4000 warmup steps. All the models were fine-tuned for 5 epochs with the batch size 16 on one NVIDIA RTX 2080Ti GPU. We selected the checkpoint for each model where the model obtains the lowest perplexity score on the Valid set.

6.3 Automatic Evaluation

The automatic evaluation uses the golden responses as reference to evaluate the responses generated by

[9] DA || EM has the same input representation except the speaker embeddings as EmpTransfo, but is instead fine-tuned from GPT-2 rather than GPT. Besides, we did not adopt the next sentence prediction (NSP) task as in (Zandie and Mahoor,

2020), because we empirically found that adding NSP leads to worse performance.

[10] https://pytorch.org/

[11] https://github.com/huggingface/transformers

	Models	PPL	B-2	R-L	Greedy
Happy	**Vanilla**	18.82	5.95*	15.00*	66.09*
	+CM	18.21	6.67*	17.64*	66.95*
	+DA	18.01	7.18*	18.09*	67.35*
	+EM	17.88	7.51*	18.27*	67.78*
	CM || DA	17.83	7.76*	18.85*	67.78*
	CM || EM	17.57	8.17*	19.58*	68.25*
	DA || EM	17.38	8.37*	19.91*	68.59*
	CM || DA || EM	17.26	9.21	20.75	68.86
	CM → DA	17.69	7.95*	18.96*	67.79*
	CM → EM	17.45	8.04*	19.49*	68.08*
	DA → EM	17.28	8.73*	20.09*	68.59*
	CM → DA → EM	**17.02**	**9.44**	**20.76**	**68.92**
	Vanilla	22.11	5.66*	13.75*	68.40*
	+CM	21.44	6.65*	17.62*	69.68*
	+DA	21.34	7.11*	17.44*	69.67*
	+EM	21.26	6.75*	17.40*	69.63*

Offmychest	**CM || DA**	21.07	7.56*	18.41*	70.16*
	CM || EM	20.83	7.78*	18.97*	70.34*
	DA || EM	20.85	7.48*	18.49*	70.19*
	CM || DA || EM	20.63	8.23	19.32	70.54
	CM → DA	20.87	7.70*	18.58*	70.33*
	CM → EM	20.72	7.71*	18.63*	70.31*
	DA → EM	20.68	7.89*	18.66*	70.25*
	CM → DA → EM	**20.35**	**8.35**	**19.54**	**70.68**

Table 2: Results of automatic evaluation. The best results are in **bold**. DA || EM is consistent with EmpTransfo (Zandie and Mahoor, 2020). CM → DA → EM is our devised model described in Section 5.1. Scores that are significantly worse than the best scores are marked with * (Student's t-test, p-value < 0.05).

models. However, when the responses are generated based on the predicted CM / DA / EM, it is not appropriate to compare the generated responses with the reference ones (Liu et al., 2016). Thus, in automatic evaluation we only considered the setting where the models are fed with the ground truth empathy factors. The results where the generated responses are based on the predicted factors will be analyzed in the later experiments.

The automatic metrics we adopted include perplexity (**PPL**), BLEU-2 (**B-2**) (Papineni et al., 2002), ROUGE-L (**R-L**) (Lin, 2004), and the BOW Embedding-based (Liu et al., 2016) **Greedy** matching score. The metrics except PPL were calculated with an NLG evaluation toolkit[12] (Sharma et al., 2017), where the generated responses were tokenized with NLTK[13] (Loper and Bird, 2002).

Results are shown in Table 2. We analyze the results from the following three perspectives:

General Performance Our model achieves the best performance on all the metrics on both domains, and most of the advantages over the competitors are statistically significant.

Impact of Empathy Factors The model performance vary from different combinations of empathy factors. First, considering more empathy fac-

[12] https://github.com/Maluuba/nlg-eval

[13] https://www.nltk.org/

tors always leads to better performance (e.g., CM → DA → EM > CM → EM > +EM > Vanilla). Second, EM brings the most gains to the model performance among the three factors. It may be because emotion is the most explicit factor that influences empathy expression (Sharma et al., 2020). In contrast, CM brings fewer gains than DA and EM. The reason may be that CM provides a high-level but coarse-grained guidance for empathetic response generation, lacking a fine-grained control like DA or EM. While the responses in the corpus of (Zhong et al., 2020) are not too long (≤ 30 words), we believe that CM plays an important role in generating longer empathetic responses, which may require the planning of multiple methanisms and more diverse usage of DA and EM.

Impact of Hierarchical modeling We noticed that for almost all the models that adopt multiple empathy factors, hierarchical modeling always leads to better performanc (e.g., CM → DA → EM > CM || DA || EM, DA → EM > DA || EM). This phenomenon is not trivial because the models with or without hierarchical modeling are all fed with the same empathy factors as the reference responses. It confirms our conjecture in Section 5.2 that hierarchical modeling can establish the connections between the embeddings of different factors, thus leading to a better capacity of empathy modeling. However, (CM, EM) is an exception. It may be due to that the pair (CM, EM) has a weaker correlation (the lowest manual information, Section 4.5) than other pairs.

6.4 Manual Evaluation

In manual evaluation, the models generate responses based on the empathy factors sampled from the predicted probability distributions. When sampling DA or EM, we used the Top-p filtering with $p = 0.9$ (Holtzman et al., 2019) to ensure the validness of the sampled results.

The manual evaluation is based on pair-wise comparison, and the metrics for manual evaluation include: **Fluency** (which response has better fluency and readability), **Coherence** (which response has better coherence and higher relevance to the context), and **Empathy** (which response shows bet-

Comparisons	Metrics	Win	Lose	κ
CM → DA → EM	Flu	33.3	34.8	0.330
vs.	Coh	35.3	39.3	0.431
DA → EM	Emp*	**39.3**	32.3	0.402
CM → DA → EM	Flu	37.3	34.5	0.383
vs.	Coh*	**41.6**	33.4	0.412
CM || DA || EM	Emp	43.4	39.6	0.416
DA → EM	Flu	36.2	38.5	0.381
vs.	Coh	**40.0**	35.7	0.523
DA || EM	Emp	44.7	42.0	0.497

Table 3: Results of manual evaluation. Ties are not shown. The metrics with significant gaps are marked with * (sign test, p-value < 0.05). κ denotes Fleiss' Kappa, whose values indicate fair agreement ($0.2 < \kappa < 0.4$) or moderate agreement ($0.4 < \kappa < 0.6$).

X, Y	Acc of X	Prop.	Hits@1/3 of Y

Happy	**CM || DA**	69.5	68.9	46.1*	81.5*
	CM → DA	70.2		**49.5**	**85.1**
	CM || EM	69.5	68.9	42.3	80.1*
	CM → EM	70.4		**42.8**	**82.7**
	DA || EM	40.1	34.6	50.3*	86.5*
	DA → EM	40.0		**53.5**	**89.7**
Offmychest	**CM || DA**	48.4	45.2	41.3*	67.9*
	CM → DA	49.2		**45.9**	**75.1**
	CM || EM	45.7	42.9	47.2*	74.2*
	CM → EM	46.1		**50.3**	**77.2**
	DA || EM	35.0	30.7	60.5*	84.8*
	DA → EM	34.9		**70.2**	**88.3**

Table 4: Results of the Hits@1/3 of predicting Y given that X is predicted rightly. "Prop." denotes the proportion of the cases where both models $X \,||\, Y$ and $X \rightarrow Y$ predict X rightly. Scores that are significantly improved after using hierarchical modeling are marked with * (sign test, p-value < 0.001).

ter understanding of the partner's experiences and feelings, and which response expresses empathy in the way that the annotators prefer). The pair-wise comparison is conducted between three pairs of models: (1) CM → DA → EM vs. DA → EM, (2) CM → DA → EM vs. CM || DA || EM, and (3) DA → EM vs. DA || EM. We randomly sampled 100 conversations from each test set of two domains (totally 200), and recruited three workers from Amazon Mechanical Turk for annotation.

Results are shown in Table 3. From all the three pairs, we find that the responses generated by these GPT-2-based models have similar fluency. The results of (1) indicate that further considering CM can significantly improve the empathy of generated responses, while the coherence may slightly decrease.

	Models	**CM**	**DA**	**EM**
Happy	**CM || DA**	69.6*	76.2*	-
	CM → DA	**79.3**	**83.6**	-
	CM || EM	73.8*	-	78.0*
	CM → EM	**76.6**	-	**82.4**
	DA || EM	-	77.5*	75.0*

	DA → EM	-	**87.3**	**85.7**
	CM || DA || EM	68.5*	70.3*	71.9*
	CM → DA → EM	**76.7**	**83.7**	**81.2**
Offmychest	**CM || DA**	61.8*	65.6*	-
	CM → DA	**71.4**	**74.8**	-
	CM || EM	65.4*	-	66.1*
	CM → EM	**71.1**	-	**74.6**
	DA || EM	-	63.7*	58.3*
	DA → EM	-	**79.5**	**75.1**
	CM || DA || EM	59.0*	60.8*	58.9*
	CM → DA → EM	**70.7**	**76.2**	**72.6**

Table 5: Realization scores. All the scores are significantly improved after using hierarchical modeling (sign test, p-value < 0.00001).

It may be because that the communication mechanisms like interpretation sometimes lead to the responses that are less relevant to the contexts (especially those sharing experiences). The results of (2) and (3) indicate that hierarchical modeling improves the coherence of generated responses. The more empathy factors are modeled, the larger improvement can be obtained.

6.5 Further Analysis of Hierarchical modeling

To give further insights of the superiority of hierarchical modeling, we analyzed (1) the prediction and (2) the realization of empathy factors.

Prediction For each pair (X, Y) in (CM, DA), (CM, EM), (DA, EM), we paired the models $X \,||\, Y$ and $X \rightarrow Y$ for comparison. Our purpose is to observe whether the prediction of X improves that of Y after using hierarchical modeling. Note that when taking the ground truth as reference, it is not appropriate to directly judge the prediction accuracy by comparing $\widehat{Y}$ and Y^* if $\widehat{X} \neq X^*$. We thus computed the conditional probability that Y is predicted rightly given that X is predicted rightly: $\mathbb{P}\left(\widehat{Y} = Y^* \middle| \widehat{X} = X^*\right)$.

Results are shown in Table 4. While the accuracy of predicting X of $X \,||\, Y$ and $X \rightarrow Y$ is close, the prediction of Y is significantly enhanced by hierarchical modeling. The results demonstrate that hierarchical modeling enables the model to select more proper empathy factors.

Realization Recall that in manual evaluation, the models generate a response based on the sampled empathy factors $\widehat{C}_y, \widehat{A}_y, \widehat{E}_y$. To verify whether these factors are well realized, we used the classifiers in Section 4.3 to identify the empathy factors displayed in the generated responses. Suppose that the identification results are $\widetilde{Z}_y, \forall Z \in \{C, A, E\}$, we computed the ratio of $\widehat{Z}_y = \widetilde{Z}_y$ as the realization score of Z.

Results are shown in Table 5. The realization of all the factors is significantly improved by hierarchical modeling. It is intuitive because hierarchical modeling can avoid the cases where the sampled factors are inappropriate or even conflicting, thus reducing the noise of empathy factors in response generation.

6.6 Case Study

We show the generated responses with different empathy factors in Figure 4. The adoption of the CM *emotional reaction* causes our model to express the same EM *admiration* (*i'm proud of you!*) as DA → EM (*good for you, man!*), while the two models generate the same sentence (*keep it up!*) when taking the DA *encouraging*. However, the further adoption of the CM *interpretation* causes our model to further share its own experiences and feelings (*i have been sober for about 10 years, and it's the best feeling ever*). As a result, with the enhancement of multiple empathy factors, the response generated by our model is more engaging and empathetic while maintaining the coherence.

Besides, we noticed another phenomenon occurring when all the three CMs are adopted. In this case, the three CMs are usually represented separately in different sentences (e.g., *I am so happy for you! I also had tried to be sober but failed. How did you make it?*), which is consistent with the results of empathetic rewriting (Sharma et al., 2021). Recall that we add the same CoMAE embeddings for all the tokens in the response during generation (Section 5.1). Such uniform operation seems non-optimal for the nonuniform realization of different CMs, especially when generating a longer empathetic response that contains multiple sentences with different CMs, DAs or EMs. We believe there is still much room of improvement when applying our CoMAE framework to longer response generation, like combining CoMAE's multi-factor hierarchical modeling with planning-based

Post you might remember me posting here when i had less than a month sober a little while back. well, yesterday i hit 100 days without alcohol and celebrated by solo hiking my state's tallest mountain!
Golden ok that is an awesome pic! love it and the story thank you!
EM you're doing great!
DA → EM good for you, man! keep it up!

CM → DA → EM (Ours)
i'm proud of you! i have been sober for about 10 years, and it's the best feeling ever. keep it up!

Figure 4: Responses generated with different empathy factors. All the generated responses express the EM *admiration*. **DA → EM** takes the DA *encouraging*. **Ours** further adopts the CM *emotional reaction* and *interpretation*.

dialog generation methods (Ghazarian et al., 2021).

7 Conclusion

In this paper, we present a multi-factor hierarchical framework CoMAE for empathetic response generation. It contains three key factors of empathy expression: communication mechanism, dialog act and emotion, and models these factors in a hierarchical way. With our devised CoMAE-based model, we empirically demonstrate the effectiveness of these empathy factors, as well as the necessity and importance of hierarchical modeling.

As future work, the CoMAE framework can be naturally extended to more factors that relate to empathy expression, such as persona (Zhong et al., 2020), by exploring the hierarchical relationships between different factors.

Acknowledgments

This work was supported by the NSFC projects (Key project with No. 61936010 and regular project with No. 61876096). This work was also supported by the Guoqiang Institute of Tsinghua University, with Grant No. 2019GQG1 and 2020GQG0005.

References

Scott Brave, Clifford Nass, and Kevin Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, 62(2):161–178.

Patrício Costa, Raquel Alves, Isabel Neto, Pedro Marvao, Miguel Portela, and Manuel Joao Costa. 2014. Associations between medical student empathy and personality: a multi-institutional study. *PloS one*, 9(3):e89254.

Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy. *Journal of Personality and Social Psychology*.

Frederique De Vignemont and Tania Singer. 2006. The empathic brain: how, when and why? *Trends in cognitive sciences*, 10(10):435–441.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4):399.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly

Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Sarik Ghazarian, Zixi Liu, Tuhin Chakrabarty, Xuezhe Ma, Aram Galstyan, and Nanyun Peng. 2021. Discol: Toward engaging dialogue systems through conversational line guided response generation. *arXiv preprint arXiv:2102.02191*.

Ursula Hess and Agneta Fischer. 2014. Emotional mimicry: Why and when we mimic emotions. *Social and Personality Psychology Compass*, 8(2):45–57.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Carl Jung. 2016. *Psychological types*. Routledge.

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

Jonathan Tarter Klein. 1998. *Computer response to user frustration*. Ph.D. thesis, Massachusetts Institute of Technology.

Mark R Leary and Ashley Batts Allen. 2011. Personality and persona: Personality processes in self-presentation. *Journal of personality*, 79(6):1191–1218.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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.

K Liu and Rosalind W Picard. 2005. Embedded empathy in continuous, interactive health assessment. In *CHI Workshop on HCI Challenges in Health Assessment*, volume 1, page 3. Citeseer.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*.

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)*, pages 8968–8979, Online. Association for Computational Linguistics.

Becky Lynn Omdahl. 2014. *Cognitive appraisal, emotion, and empathy*. Psychology Press.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: a survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):1–40.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037. Curran Associates, Inc.

Martin Popel and Ondřej Bojar. 2018. Training tips for the transformer model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Hannah Read. 2019. A typology of empathy and its many moral forms. *Philosophy Compass*, 14(10):e12623.

Nadine R Richendoller and James B Weaver III. 1994. Exploring the links between personality and empathic response style. *Personality and individual Differences*, 17(3):303–311.

Babette Rothschild. 2006. *Help for the helper: The psychophysiology of compassion fatigue and vicarious trauma*. WW Norton & Company.

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 *The World Wide Web Conference*.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Con-*

ference on Empirical Methods in Natural Language Processing (EMNLP), pages 5263–5276, Online. Association for Computational Linguistics.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL 2021*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rohola Zandie and Mohammad H Mahoor. 2020. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *arXiv preprint arXiv:2003.02958*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566, Online. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018b. The design and implementation of xiaoice, an empathetic social chatbot. *arXiv preprint arXiv:1812.08989*.

A Emotion Mapping

In the original paper of (Demszky et al., 2020)[14], the authors provide the hierarchical clustering results of the 27 emotions (Figure 2 in their paper), which reflect the nested structure of their proposed emotion taxonomy. Based on the clustering results, we merged the emotions that are highly correlated with each other, and the mapping between our adopted emotions and the original emotions is shown in Table 6.

Ours	**Original**
admiration	admiration, pride
anger	anger, annoyance, disgust, disapproval
approval	approval, realization
caring	caring, desire, optimism
fear	fear, nervousness
gratitude	gratitude, relief

joy	joy, amusement, excitement, love
sadness	sadness, disappointment, embarrassment, grief, remorse
surprise	surprise, confusion, curiosity

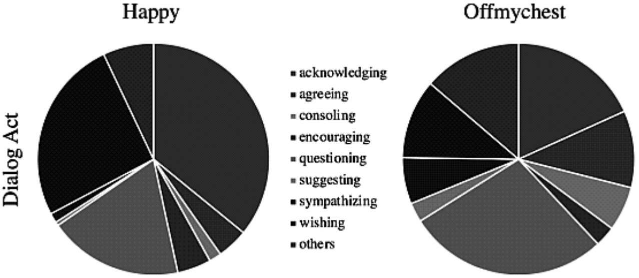
Table 6: Mapping between our adopted emotions and the original emotions in (Demszky et al., 2020).

B Statistics of Annotation

We computed the proportions of the last responses annotated with ER / IP / EX. In the Happy domain, the proportions are 76.0% / 10.2% / 18.7%, while in the Offmychest domain are 57.1% / 21.4% / 27.9% respectively. The statistics of DA and EM are shown in Figure 5.

We can find several differences between two domains. In terms of **communication mechanism**, the responses in the Offmychest domain prefer *interpretation* and *exploration*, while *emotional reaction* occupies a larger proportion in the Happy domain. In terms of **DA**, the actions that provide support (such as *agreeing*, *consoling*, *suggesting*, and *sympathizing*) are more frequently adopted in the Offmychest domain. It is similar when it comes to **emotion**, where the emotions such as *approval* and *caring* are displayed more commonly when responding to the posts with negative sentiments. We also observed that the responses in the Offmychest domain may also display the emotions like *anger* and *sadness*, indicating that they do understand

[14] `https://arxiv.org/abs/2005.00547v2`



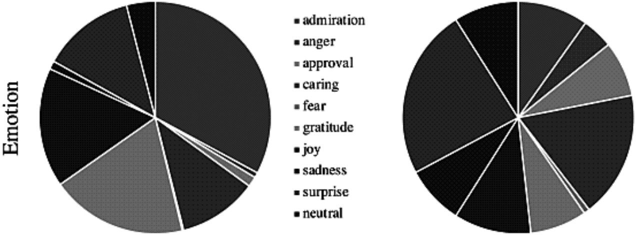


Figure 5: Statistics of the annotation results of DA and EM on the two domains.

the experiences and feelings of the conversation partners.