Emstremo: Adapting Emotional Support Response with Enhanced Emotion-Strategy Integrated Selection

Junlin Li, Bo Peng, Yu-Yin Hsu
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
11 Yuk Choi Rd, Hung Hom, Hong Kong SAR
junlin.li@connect.polyu.hk, bopeng@polyu.edu.hk, yyhsu@polyu.edu.hk

Abstract

To provide effective support, it is essential for a skilled supporter to emotionally resonate with the help-seeker's current emotional state. In conversational interactions, this emotional alignment is further influenced by the comforting strategies employed by the supporter. Different strategies guide the interlocutors to align their emotions in nuanced patterns. However, the incorporation of strategy into emotional alignment in the context of emotional support agents remains underexplored. To address this limitation, we propose an improved emotional support agent called Emstremo. Emstremo aims to achieve strategic control of emotional alignment by perceiving and responding to the user's emotions. Our system's state-of-the-art performance emphasizes the importance of integrating emotions and strategies in modeling conversations that provide emotional support. (The code for Emstremo is available at https://github.com/CN-Eyetk/Emstremo)

Keywords: emotional support conversation, emotion-strategy integration, emotion state, support-strategy, response generation

1. Introduction

Emotional support brings substantial benefits to friendships, relationships, health, and overall well-being (Fehr, 2004; Uchino et al., 1996; Cohen and Wills, 1985). However, providing effective emotional support can be challenging and elusive. Support that lacks vicarious emotion or empathy may unintentionally result in negative outcomes (Chen and Xu, 2021; Holmstrom et al., 2005). For instance, responding in an emotionally 'cold' manner (e.g., Response I in Figure 1) or criticizing and dismissing the recipient's feelings hinders the achievement of emotional support as it fails to align with the recipient's emotional state (Spottswood et al., 2013). Another challenge arises from the inappropriate use of comforting strategies (e.g., Response II in Figure 1), which can undermine the perceived support from the help-seeker's perspective (Burleson, 2003). For example, giving advice without deliberately considering the situation or dialogic context (Goldsmith and Fitch, 1997), especially in the early stages of communication, can be risky. Therefore, a reliable emotional support model should possess skillful decision-making abilities in both emotions and strategies.

It is important to note that strategy and emotion are not totally independent of each other in the context of emotional support conversations. Emotional support conversation also entails a dynamic interplay between the affective and behavioral factors (Scarantino, 2018; Saha et al., 2021). Placing emotional support in a conversation context, the choice of support strategy plays a pivotal role in modulating how the supporter aligns their emotions with the recipient's feelings. In some self-oriented strategies, such as sharing similar experiences or engaging in self-disclosure, the supporter tends to express similar emotions with the seeker to maintain empathy or sympathy (Meng and Dai, 2021) (See Response IV in Figure 1). In other-oriented strategies (e.g., reflecting feelings, assuring) or dialogic strategies (e.g., questioning), however, the supporter may not need to mirror the seeker's emotion in a one-to-one manner (Burleson, 2008, 2003) (See Response V and VI in Figure 1). Noticing the decisive role of strategy in shaping vicarious emotion, it is recommended to explicitly integrate the emotion-strategy interplay into emotional support agents.

Previous attempts at Emotional Support Conversation (ESC) have been continuously emphasizing the perception of seeker (user) emotion (Zhao et al., 2023; Cheng et al., 2022; Zhou et al., 2023). However, it is still unanswered how to predict or tailor the production of supporter (system) emotion to the user's feelings in the territory of ESC. Besides, the integration between strategy and emotional alignment has never been given attention across existing methods in ESC or the scope of emotional conversation. To cope with such limitation, we introduce a novel approach called **Emstremo**, which improves the empathy of generated responses by sensitively controlling verbal emotions in response to diverse seeker emotions while dynamically tailored with appropriate support strategies.

Building on the aforementioned insights, we aim to enhance response generation in ESC by incorporating a sensitive response emotion control. To achieve this goal, we propose the use of a strategy-adapted emotion transition matrix to model the

LREC-COLING 2024, pages 5794–5805
20-25 May, 2024.

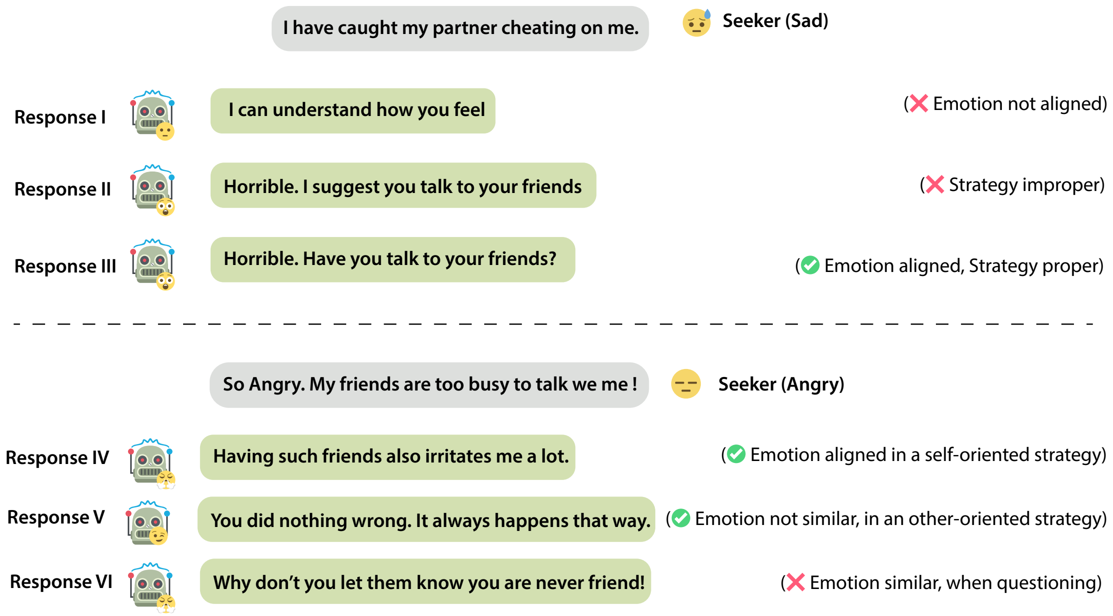
I have caught my partner cheating on me.
Seeker (Sad)
Response I
I can understand how you feel
(✗ Emotion not aligned)
Response II
Horrible. I suggest you talk to your friends
(✗ Strategy improper)
Response III
Horrible. Have you talk to your friends?
(✓ Emotion aligned, Strategy proper)
So Angry. My friends are too busy to talk we me !
Seeker (Angry)
Response IV
Having such friends also irritates me a lot.
(✓ Emotion aligned in a self-oriented strategy)
Response V
You did nothing wrong. It always happens that way.
(✓ Emotion not similar, in an other-oriented strategy)
Response VI
Why don't you let them know you are never friend!
(✗ Emotion similar, when questioning)

Figure 1: An example of emotional support conversation. We expect an empathetic ESC system to select the appropriate strategy, and carefully control its verbal emotion according to the seeker's emotion and the strategy it employs.

emotion-to-emotion alignment adjusted to different support strategies, by which we integrate a three-way interaction into the emotional control over response generation. By employing this emotion control mechanism, our approach successfully improved the empathy of the generated responses and outperformed most previous methods .

Our work makes two major contributions.

- We enhance the ESC system by incorporating an interactive emotion alignment mechanism. This mechanism predicts and controls the appropriate response emotion while taking into account both the seeker's emotional state and the supporter's strategy.
- Our second major contribution lies in the improvement of the commonsense-aware emotion support system, specifically in terms of the similarity to the ground truth and the diversity. These enhancements are validated by human evaluations, which demonstrate the notable advancement of our model in terms of its ability to exhibit empathy.

2. Related Work

2.1. Emotional Support Conversation

Emotional Support Conversation (ESC) System is a growing field, particularly with the development of the emotional support conversation dataset (ESCONV) (Liu et al., 2021). Compared with similar tasks such as empathetic or emotional conversation (Rashkin et al., 2018; Li et al., 2017), ESC involves a unique scenario where the system is required to choose appropriate support strategies based on the dialogue context, seeker's situation, and emotional state (Liu et al., 2021).

In ESC, the state-of-the-art has incorporated commonsense knowledge extracted from COMET-ATOMIC (Hwang et al., 2021) as shown in most existing ESC systems (Tu et al., 2022; Peng et al., 2022; Zhao et al., 2023; Deng et al., 2023; Zhou et al., 2023). Moreover, graph networks have been extensively used to capture the turn-to-turn transition within the dialogue context as well as between dialogue and situational factors (Peng et al., 2022; Zhao et al., 2023; Zhou et al., 2023), with a particular focus on modeling emotion transitions in the dialogue context (Zhao et al., 2023). However, we argue that the emotional transition between context and upcoming utterance is equally a key ingredient of an effective ESC system.

2.2. Dialogue Strategy Selection

Effective strategy identification and control are crucial aspects of an ESC system. By prepending special tokens, Liu et al. (2021) conditions response generation on the selection of a comforting strategy. Tu et al. (2022) propose cross-attention networks to integrate commonsense knowledge, which enhances the identification of strategy, and conditions response generation on mixed strategies. Zhao et al. (2023) uses a graph network to model turn-level semantic, emotion, and strategy state transitions within the dialogue history, which significantly enhances the identification of strategy. Recently, seeker feedback has been leveraged to enhance the emotional support dialog system. Cheng et al. (2022) employ data augmentation to train a feedback predictor, which directs the look-ahead planning during strategy selection. Peng et al. (2023) utilizes turn-level and conversation-level feedback to encourage the appropriate decision of support strategy.

However, most of the aforementioned attempts often overlook the role of supportive strategy as a potential modulator of the emotional transition between seeker and supporter. As we have highlighted above, the choice of strategy explicitly determines how the seeker's emotion impacts the supporter's response emotion, ultimately leading to an optimal emotional support outcome (Meng and Dai, 2021; Burleson, 2008, 2003).

2.3. Emotional Response Generation

Performing appropriate or sensitive emotion in response generation has been a central concept in emotional or empathetic dialogue systems. Rashkin et al. (2018) uses emotion-wise special tokens to condition the verbal emotion of the generated response. Wang et al. (2022) uses an emotional intent selection module in the proposed empathetic dialogue system. Using a variational encoder, Majumder et al. (2020) conditioned the response emotion on the emotional valence of the user's emotion state. Ma et al. (2024) incorporates personality into the generation of system emotion. Irfan et al. (2020) improves the emotion transition in the conversation using user feedback. However, the relationship between dialogue acts (such as support strategies) and response emotion has often been overlooked (Li et al., 2019, 2022b; Sabour et al., 2022; Cai et al., 2023; Zhou et al., 2022).

Recently, response emotion control has received a lot of attention in ESC. Zhou et al. (2023) leverages reinforcement learning to improve the emotional positivity of generated response. However, we argue that emotional positivity should not be imposed as a constant “pressure" of response emotion control, because certain strategies, such as self-disclosure, can alleviate the “pressure" (Meng and Dai, 2021). Therefore, response emotion control should be linked to strategy selection as a dialogue act factor.

3. Methodology

Fig. 2 displays an overview of the architecture of our model. This model contains three major components, including a dialogue encoder, a strategy-adapted emotion aligner, and a decoder of emotion-strategy-controlled response.

3.1. Preliminaries

We can formulate a minimalist ESC problem as: Given dialogue history $X = (u_1, u_2, \cdots, u_T)$ and situation description s, to maximize $p(Y|X, s)$.

To achieve strategy-adapted emotional alignment, we enrich this problem with emotion identification, strategy decision, and emotion prediction: Given X, to predict user emotion e, strategy g, and system emotion v. Conditioning the system response on the affective and strategic factors, we reformulate the response generation probability $p(Y|X, s, v, g)$.

3.2. Dialogue Encoder

We encode the dialogue as $\mathbf{X} = \mathrm{Encoder}(\mathtt{CLS}, u_1, \mathtt{EOS}, u_2, \ldots, u_T, \mathtt{EOS})$. where CLS and EOS are the start and separation tokens bewteen two utterances.

To enable self-other differentiation, we initialize two speaker embeddings $E^r \in \{\mathbf{e}^{usr}, \mathbf{e}^{sys}\}$ ($\mathbf{e}^{usr} \in \mathbb{R}^d$ and $\mathbf{e}^{sys} \in \mathbb{R}^d$). At the embedding layer, we reach token embedding by fusing its word embedding E^w with the role embedding of the current speaker E^r, through a fully connected network $\mathrm{FC}_{\text{fusion}}$.

$$E^u = \mathrm{FC}_{\text{fusion}}(E^w \oplus E^r) \tag{1}$$

we feed token embeddings into the self-attention block to get $\mathbf{X}$ as the hidden states of each token.

3.3. Strategy-adapted Emotion Aligner

Taking the hidden state of CLS as the dialogue hidden state $\mathbf{h} = \mathbf{X}[0]$, we identify user emotion $e \in \{e_1, \cdots, e_c\}$ and system strategy $g \in \{g_1, \cdots, g_n\}$, from the interaction of which we derive system emotion $v \in \{v_1, \cdots, v_m\}$.

Initially, we use trainable parameters W^e and W^g to infer user emotion distribution and system strategy policy.

$$\mathbf{p}^e = \mathrm{softmax}(W^e(\mathbf{h})) \tag{2}$$

$$\mathbf{p}^g = \mathrm{softmax}(W^g(\mathbf{h})) \tag{3}$$

Suppose we have c as the dimension of user emotion space and n as the dimension of strategy space, we further initialize and train n matrices $\{\hat{\mathbf{T}}_1, \ldots, \hat{\mathbf{T}}_n\} \in \mathbb{R}^{m \times c}$ to convert user emotion distribution $\mathbf{p}^e$ to system emotion distribution $\mathbf{p}^v$ in each strategy. To reach the expected system emotion, we pool each strategy-specific emotion distribution over the strategy policy.

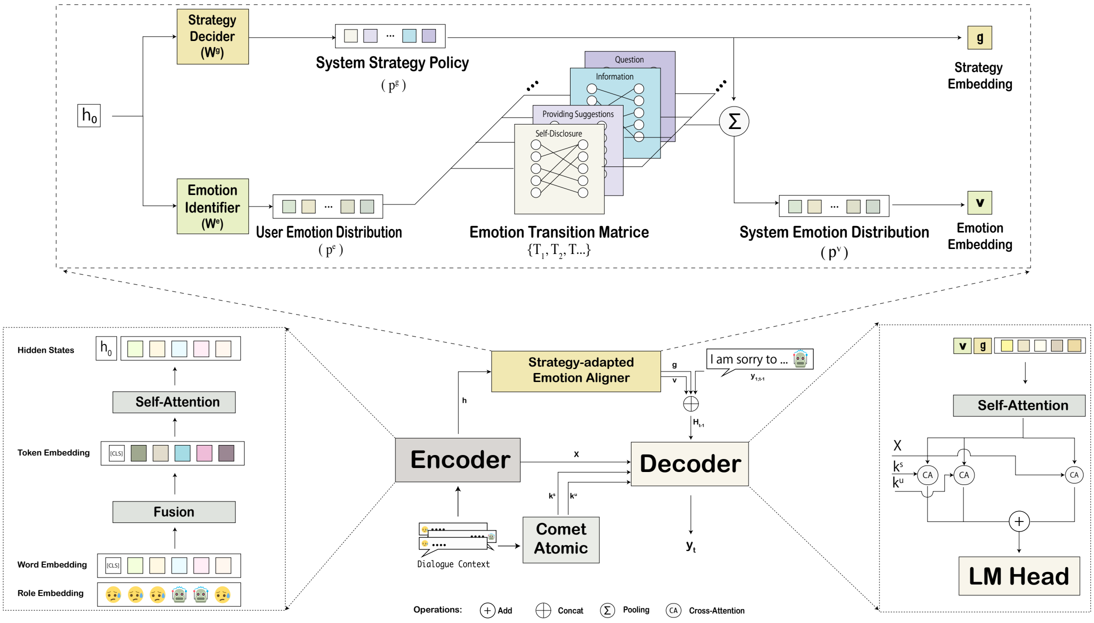

Figure 2: The overview of Emstremo

$$\mathbf{p}^v = \sum_{i=1}^{n} \mathbf{p}^g(g = g_i) \cdot \mathbf{T}_i\left(\mathbf{p}^e\right) \tag{4}$$

To reduce the distance between $\mathbf{p}^v$ and the emotion distribution of ground truth response $\tilde{\mathbf{p}}^v$, we use a pre-trained emotion classifier to represent the ground-truth emotion as a distribution $\tilde{\mathbf{p}}^v$ from the ground-truth response Y, and use KL divergence to supervise the distance between $\mathbf{p}^v$ and $\tilde{\mathbf{p}}^v$ during training.

$$\tilde{\mathbf{p}}^v = \text{Classifier}(Y) \tag{5}$$

$$\mathcal{L}_v = D_{KL}\left[\mathbf{p}^v \mid\mid \tilde{\mathbf{p}}^v\right] \tag{6}$$

3.4. Strategy-controlled Emotional Response Generator

Inspired by the prefix-tuning technique (Li and Liang, 2021), we prepend strategy parameters $\mathbf{g}$ and emotion parameters $\mathbf{v}$ to the beginning of the embedded tokens, by which we control the generation process strategically and effectively. The prefixed parameters are reached by weighting a set of trainable strategy embeddings $\{\mathbf{e}^g_1, \ldots, \mathbf{e}^g_n\}$ and a set of emotion embeddings $\{\mathbf{e}^v_1, \ldots, \mathbf{e}^v_m\}$ respectively over $\mathbf{p}^g$ and $\mathbf{p}^v$.

$$\mathbf{g} = \sum_{i=1}^{n} \mathbf{p}^g(g = g_i) \cdot \mathbf{e}^g_i \tag{7}$$

$$\mathbf{v} = \sum_{i=1}^{m} \mathbf{p}^v(v = v_i) \cdot \mathbf{e}^v_i \tag{8}$$

Where $\mathbf{e}^g_i \in \mathbb{R}^d$ and $\mathbf{e}^v_i \in \mathbb{R}^d$.

Suppose the embedding of decoder input Y_{t-1} is $\mathbf{E}_{t-1}$, we treat $\mathbf{H}_{t-1} = \mathbf{v} \oplus \mathbf{g} \oplus \mathbf{E}_{t-1}$ as input of Q and V projection.

We use cross-attention to model the correlation with the context $\mathbf{X}$.

$$\mathbf{Y} = \text{CrossAttn}\left(\mathbf{H}_{t-1}, \mathbf{X}\right) \tag{9}$$

In addition, we enrich the decoding procedure with COMET-ATOMIC commonsense knowledge. Following Tu et al. (2022)'s operation, we infer and encode two sets of knowledge, (1) the top-N tail nodes $\{K^s_1, \cdots, K^s_N\}$ based on situation description s, and (2) the top-M tail nodes $\{K^u_1, \cdots, K^u_M\}$ based on the user's last post u_T. Using the dialogue encoder, we represent each node as a hidden state, $\{\mathbf{k}^s_1, \cdots, \mathbf{k}^s_N\}$ for knowledge derived from s and $\{\mathbf{k}^u_1, \cdots, \mathbf{k}^u_M\}$ for knowledge derived from u_T, and concat each of the two set of knowledge on the dimension of node (namely timestep dimension for dialogue encoder). More details are available in Appendix A.2.

$$\mathbf{k} = \text{Encoder}(K)\left[0\right] \tag{10}$$

$$\mathbf{H}^s = \bigoplus_{i}^{N} \mathbf{k}^s_i \tag{11}$$

$$\mathbf{H}^u = \bigoplus_{i}^{M} \mathbf{k}^u_i \tag{12}$$

We devise two other cross-attentions to model the correlation with situation-derived and utterance-derived commonsense knowledge. Then, we aggregate the output of three cross-attentions as the output of each layer.

$$\mathbf{Y}^s = \mathrm{CrossAttn}\left(\mathbf{H_{t-1}}, \mathbf{H}^s\right) \tag{13}$$

$$\mathbf{Y}^u = \mathrm{CrossAttn}\left(\mathbf{H_{t-1}}, \mathbf{H}^u\right) \tag{14}$$

$$\hat{\mathbf{Y}} = \mathbf{Y} + \mathbf{Y}^s + \mathbf{Y}^u \tag{15}$$

Suppose the hidden state of the final token is $\hat{\mathbf{y}}_{t-1}$, we predict the next token from the language model head and calculate the LM loss.

$$\mathbf{p}(y_t|X, s, \mathbf{g}, \mathbf{v}) = \mathrm{softmax}(\mathrm{FC}_{\mathrm{Dec}}(\hat{\mathbf{y}}_{t-1})) \tag{16}$$

$$\mathcal{L}_{gen} = -\log \mathbf{p}(Y|X, s, \mathbf{g}, \mathbf{v}) \tag{17}$$

To guarantee the effectiveness of strategy control where our system centers on, we introduce a contrastive learning loss to expand the distance between utterances produced under different strategies. During training, we impose contrastive learning loss over the hidden state of the last token $\mathbf{y}^{eos}$, according to the ground truth strategy label $g*$.

$$\begin{aligned} \mathcal{L}_{cont} &= -\mathbf{1}\left[g_i^* = g_j^*\right] f\left(\mathbf{y}_i^{eos}, \mathbf{y}_j^{eos}\right) \\ &-\mathbf{1}\left[g_i^* \neq g_j^*\right] \max\left(0, \epsilon - f\left(\mathbf{y}_i^{eos}, \mathbf{y}_j^{eos}\right)\right) \end{aligned} \tag{18}$$

Where f is the cosine-similarity function, and $\epsilon = 2.0$.

We compute seeker emotion loss and strategy selection loss based on the ground truth seeker emotion label e^* and seeker strategy label g^*.

$$\mathcal{L}_e = -\log \mathbf{p}(e = e^*) \tag{19}$$

$$\mathcal{L}_g = -\log \mathbf{p}(g = g^*) \tag{20}$$

The final learning objective is defined as a combination of generation loss, classification loss, contrastive loss, and the divergence between ground truth and predicted supporter emotion distribution:

$$\mathcal{L} = \mathcal{L}_{gen} + \alpha\mathcal{L}_{cont} + \gamma\left(\mathcal{L}_e + \mathcal{L}_g + \mathcal{L}_v\right) \tag{21}$$

4. Experiments

4.1. ESConv Dataset

Our experiments use the Emotional Support Conversation Dataset, ESConv (Liu et al., 2021). The dataset contains 1,300 long dialogues (29.5 utterances on average) with 38,350 utterances (16.7 tokens on average). Each dialogue is annotated with an emotion label representing the emotional state of the seeker. The annotation of strategy adopts an 8-label set [1].

Regarding dataset preprocessing, two divisions have been commonly adopted across previous studies. Some have embraced the official division (e.g. Cheng et al. (2022); Peng et al. (2023); Deng et al. (2023)). Others follow an “8-1-1" division (e.g., Tu et al. (2022); Zhao et al. (2023); Peng et al. (2023)) introduced in Tu et al. (2022). We refer to the former as “Official Division" and the latter as “MISC Division". More statistics are available in appendix A.1.

4.2. Baselines

We compare our model with five blenderbot-based systems including **BlenderBot-Joint** (Liu et al., 2021), **TransESC** (Zhao et al., 2023), **MISC** (Tu et al., 2022), **KEMI** (Deng et al., 2023), **FADO** (Peng et al., 2023), and one bart-based system which is **MultiESC** (Cheng et al., 2022). We reproduce **TransESC** (Zhao et al., 2023), **KEMI** (Deng et al., 2023), and **MISC** (Tu et al., 2022) based on their official repositories. ,

4.3. Implementation Details

We finetune `facebook/blenderbot_small-90M` (Roller et al., 2020) and `facebook/bart-base` (Lewis et al., 2019) on an Nvidia Geforce Rtx 3090 GPU. The AdamW optimizer (Loshchilov and Hutter, 2017) is set to train our model with β_1 = 0.9 and β_2 = 0.999. The batch size of training is 20. We control the learning rate during the training process with an initial learning rate of 2e-5 and a linear warmup with 120 warmup steps. Consistent with the comparable SOTA systems, we adopt the decoding algorithms of Top-p and Top-k sampling with p=0.3, k=30, temperature =0.7 and the repetition penalty 1.03 (Zhao et al., 2023; Tu et al., 2022). For the loss function (Equation 21), the α was set to 0.2, and γ was set to 0.05.

The emotion classifier adopted in Equation 5 is `SamLowe/roberta-base-go_emotions`, which is trained on a fine-grained emotion classification dataset “Go-Emotions" (covering 28 emotion categories) (Demszky et al., 2020), based on `roberta-base` model (Liu et al., 2019). To acquire commonsense knowledge from the user’s last utterance and the user’s situation description, we use all nine relations available in `COMET-ATOMIC` (Bosselut et al., 2019), and set N=30 (Equation 11) and M=20 (Equation 12). In practice, the knowledge encoder shares weights with the diaogue encoder.

We train the `blenderbot-small` model for 8 epochs and `bart-base` model for 10 epochs. The saving step was set to 300. To ensure a fair comparison, we take the check-point with the lowest perplexity to evaluate.

[1] The 8 categories are [Questions], [Self-disclosure], [Affirmation and Reassurance],[Providing Suggestions], [Other], [Reflection of feelings], [Information], [Restatement or Paraphrasing]

	B-1(↑)	**B-2**(↑)	**B-3**(↑)	**B-4**(↑)	**MET**(↑)	**R-L**(↑)	**D-1**(↑)	**D-2**(↑)
Bart-based (Official Division)								
MultiESC (Cheng et al., 2022)	21.65	9.18	4.99	**3.09**	8.84	**20.41**	-	-
Emstremo (Bart)	**23.43**	**9.87**	**5.16**	3.05	**9.23**	19.98	2.94	14.13
Blenderbot-based (Official Division)								
BlenderBot-Joint (Liu et al., 2021)	17.08	5.52	2.16	1.29	-	15.51	2.71	19.38
KEMI (Deng et al., 2023)	20.76	8.51	4.38	2.54	8.12	17.30	3.01	15.79
FADO (Peng et al., 2023)	-	8.00	4.00	2.32	-	17.53	**3.84**	**21.84**
Emstremo (BlenderBot)	**20.96**	**8.80**	**4.59**	**2.75**	**8.66**	**20.48**	2.90	14.80
Blenderbot-based (MISC Division)								
BlenderBot-Joint (Liu et al., 2021)	18.78	7.02	3.2	1.63	-	14.92	2.96	17.87
FADO (Peng et al., 2023)	-	8.31	4.36	2.66	-	18.09	3.8	**21.39**
MISC (Tu et al., 2022)	17.95	7.20	3.65	2.13	7.68	17.91	3.87	17.31
TransESC (Zhao et al., 2023)	18.58	7.62	3.91	2.31	7.88	17.92	4.05	18.60
Emstremo (BlenderBot)	**19.36**	**8.52**	**4.72**	**2.99**	**8.23**	**19.35**	**4.59**	19.66
Ablation Analysis (Official Division)								
Emstremo (BlenderBot)	**20.96**	**8.80**	**4.59**	**2.75**	**8.66**	**20.48**	2.90	**14.80**
w/o_Emotion2Response	18.87	7.89	4.12	2.45	7.79	17.78	**2.91**	14.24
w/o_Strategy2Emotion	20.88	8.61	4.43	2.64	8.26	18.18	2.74	13.90
w/o_Strategy2Response	20.05	8.33	4.35	2.61	8.09	17.81	2.79	14.05

Table 1: Automatic evaluation results of our model *Emstremo*, compared against the state-of-the-art baselines. Blender-Bot-based models and Bart-based models are separately compared. The best results among each group are highlighted in bold.

4.4. Evaluation Metrics

We adopt both automatic and human evaluations. Four sets of **automatic metrics** for evaluation are: (1) BLEU-1 (**B-1**), BLEU-3 (**B-3**), BLEU-4 (**B-4**), METEOR (**MET**), and ROUGE (**R-L**) to measure the similarity between the generated response and its ground-truth response; (2) Distinct-*n* (**D**-*n*) to capture the diversity of the generated response.

Following See et al. (2019), we recruited 4 annotators, who are Ph.D. students in linguistics and psychology, to evaluate the responses generated by 3 models (**MISC**, **TransESC**, and our **Emstremo**, trained from the MISC division) using 100 dialogue contexts from the test set of ESConv dataset as input (in MISC division). The criteria include **Fluency**, **Identification**, **Suggestion**, **Empathy** with levels of {0,1,2}. To avoid bias, the responses generated by different models are displayed in a random order. Upon evaluating a given response, the annotators had no information about which of the models was the generator of the current response. Following Zhao et al. (2023), we consider **Fluency** as referring to the coherence and readability of the responses, **Identification** as the extent to which the model explores the seeker's problems effectively, **Suggestion** as the helpfulness of the provided suggestion, and **Empathy** as the extent to which the model is empathetic in understanding the seeker's feelings and situations.

5. Results

5.1. Automatic Evaluation

As shown in Table 1, our model achieves new state-of-the-art performances in most of the generation quality metrics, especially the alignment with the ground-truth response. Compared with all the baseline models, our model achieves the best scores on BLEU-1,2,3 (**B-1,2,3**) and METEOR (**MET**).

Since METEOR score has been particularly recommended for task-oriented dialogue natural language generation (Sharma et al., 2017; Li et al., 2022a), our model's advantage on this metric demonstrates notable progress within the scope of ESC. Furthermore, it has been pointed out that the METEOR score correlates tightly with the human judgment of text similarity (Li et al., 2022a). Therefore, our model is more reliable than the baseline systems regarding the convergence toward expert response. Finally, BLEU scores have been proven as a potent indicator of dialogue coherence (Gandhe and Traum, 2008), which highlights our system's advantage in generating coherent responses.

5.2. Ablation Studies

We evaluate the impact of various processes implemented in our system. Firstly, we remove the emotion prepend from the decoder (hence **w/o_Emotion2Response**). Secondly, we remove strategy prepend from the decoder (**w/o_Strategy2Response**). Thirdly, we stop strategy from modifying the emotional alignment matrice (**w/o_Strategy2Emotion**).

Table 1 shows the performance of each ablation model in comparison to our full model (based on Blender-bot). It is evident that the optimal response quality is achieved by the integration of strategy, emotion control and strategy's modulation over emotion, since the full model outperforms all the ablation models in terms of a majority of response quality metrics. Specifically, the key contribution is attributed to **response emo-**

tion control (Emotion2Response), as its exclusion significantly undermines all the text generation metrics (See **w/o_Emotion2Response**) in Table 1. Furthermore, the exclusion of strategy control(See **w/o_Strategy2Response**) leads to relatively lower-quality responses. Finally, removing the modulating effect from strategy to emotion (See **w/o_Strategy2Emotion**) leads to a drastic reduction of “Distinct-n" metrics which is detrimental to a high-quality response generation system.

5.3. Human Evaluation

Table 2 shows the results of human ratings. Our model outperforms the two SOTA models in the **Empathy** score. This score is critical to the perceived supportiveness. Furthermore, our model yields the highest fluency, which indicates notable progress in the smoothness of response. We use the paired t-test to validate the inter-model difference of the scores given by the same annotator over the responses generated by different models from the same input. The results confirm that our method significantly enhances the empathetic capability of the automatic ESC system (*p*<0.01 for Emstremo over MISC, *p*<0.05 for over TransESC).

	MISC	**TransESC**	**Emstremo**
Fluent	1.65	1.66	**1.70**
Identification	0.95	0.97	**1.00**
Suggestion	0.81	**0.90**	0.89
Empathy	0.97	1.03	**1.13**‡

Table 2: Human Rating Scores. ‡ denotes p <0.05 in paired T-test with the other models.

5.4. Visualization Analysis

5.4.1. Inter-strategy Difference of Emotional Alignment

From **Emstremo**, we can interpret and address an appealing inquiry: How do the interlocutors emotionally resonate with each other under different strategies? To quantify the effect of “strategy shift" on “emotional alignment", we try to visualize the difference between emotion transition matrices trained for each strategy. For clarity, we focus on the “self-oriented" strategy, "Self-disclosure", and visualize its difference with other strategies in terms of the transition probability between each pair of emotions in Fig 3. Since it is highly cumbersome to visualize all the 28 emotion categories (Demszky et al., 2020), we concentrate on 5 major emotions that add up to the highest probabilities after summing up all the transition matrices.

Situation: friend has possible COVID and going to hospital today for tests
Seeker: Hello there
Supporter: Hello! How are you doing today?
Seeker:Not too bad. Been experiencing some anxiety though. It is kind of an ongoing issue for me
Supporter: Anxiety can be a real challenge. What have you done to face it in the past that works for you?
Seeker:I have tried going to support groups and therapy. I also take medications. Despite this I still feel overwhelmed.

TransESC: I understand that. i have had to deal with anxiety before.
MISC: I am sorry to hear that. do you have any friends or family that you can talk to about it?
Our: I understand. I have been in a similar situation before. I **know how difficult it can be** to get through.
Ground-truth: What do you do when it happens?

Situation: I fear disappointing my parents.
Seeker:hi, can I discuss an issue with you?
Supporter: Yes please. What has been bothering you?
Seeker:My parents, they expect me to always get straight A's.
Supporter: oh wow. Is that hard for you? Explain what's upsetting you about that.
Seeker:It is upsetting because it is an expectation. I am not perfect and I always try my hardest. I am afraid of not being perfect.

TransESC: I understand how you feel. I have been in your shoes before.
MISC: I can understand that. I have had parents who were like that, too. I think it's important to be firm with them.
Our: I understand. I have been in that position before. it's **hard to be perfect** when you're not sure what you want to do.
Ground-Truth: I think it would be a good idea for you to try and get medical help for your depression.

Situation: I work a dead end job.
Seeker:It can be difficult to express feelings. I agree.
Supporter: Do you have a specific problem that you are experiencing?
Seeker:I work a dead end job with no potential and it is depressing being there for 9 hours a day. Thats why I am on here instead of working.
Supporter: That sounds really exhausting. You work longer hours than a normal job and must feel really burned out.
Seeker:Yeah the burn out isn't the worst I experienced. I previously worked jobs that I was at for 12 hours a day but they were more rewarding.

TransESC: I think that is a good thing to have a job that you can work for.
MISC: Do you think you could find a new job?
Our: **I can imagine** that **you are feeling** a lot more **stressed** out than you think. have you thought about looking for a new job?
Ground-truth: Not bad for a Monday, but that's often a bummer of a day. What's got you down?

Table 3: Three cases of responses generated from Emstremo and other compared models

Inspecting the interesting variations between self-disclosure and other strategies, it is without dispute that the change in strategy gives rise to a noticeable shift in emotional alignment pattern. Compared with self-disclosure, we capture a more dominant role of "curiosity" in mitigating various negative feelings of support-seekers in "Question," "Restatement or Paraphrasing," and "Reflection of feelings." Such a surge of curiosity-prominence is not observed in other strategies such as "Affirmation and

Reassurance." We can also notice that "neutral" emotion is relatively inhibited in "Affirmation and Reassurance" in contrast with other strategies.

The activation or inhibition of specific emotions as a result of a "strategy shift" is further dependent on the "source emotion" or "seeker emotion" as we can see from Figure 3. For instance, curiosity is generally inhibited when shifting to " providing suggestions." However, closer inspection reveals that "curiosity" is activated when providing suggestions in the face of a support-seeker in "anger".

From the clues above, our results hint at a subtle process of emotional "re-alignment" following the change of support strategy. Such an emotional "re-alignment" addressed into different strategies fuels the enhancement of our model in response generation.

5.4.2. Emotion Change in Dialogue Flow

Previous studies in ESC field have shed intensive light on the change of strategy in dialogue flow (Liu et al., 2021; Tu et al., 2022). Owing to our effective incorporation of system emotion, this study offers sufficient insights into the change of emotion following the progress of the conversation. For brevity, Figure 4 illustrates the change of top-10 emotions' probability alongside the axis of turn order. From Figure 4, it is appealing to notice that emotions of different valance exhibit discrepant patterns in earlier and later conversations. Specifically, sadness and remorse gradually weaken with the progress of conversation, on the opposite of positive emotions such as admiration and joy. This tendency resonates with our notion that emotional support involves gradually eliciting positive emotion.

5.5. Case Studies

Table 3 offers three examples comparing the responses generated by **Emstremo** and two baselines. Overall, our responses exhibit greater levels of caring and empathy in comforting the seeker's emotional state, particularly by sharing or aligning the seeker's feelings.

Looking at the first example in Table 3, the two baseline models fall short of a caring reaction to the seeker's emotional expressions. The seeker, in the first case, has been emotionally narrating the "COVID" issues and negative feelings. However, the baseline models' response shows inadequate empathetic concern or shared feelings, as the appraisal of the seeker's difficulty is unseen. In contrast, it is our model that maps the seeker's emotion by appraising and sharing the "pain" with the seeker (i.e. "know how difficult it can be").

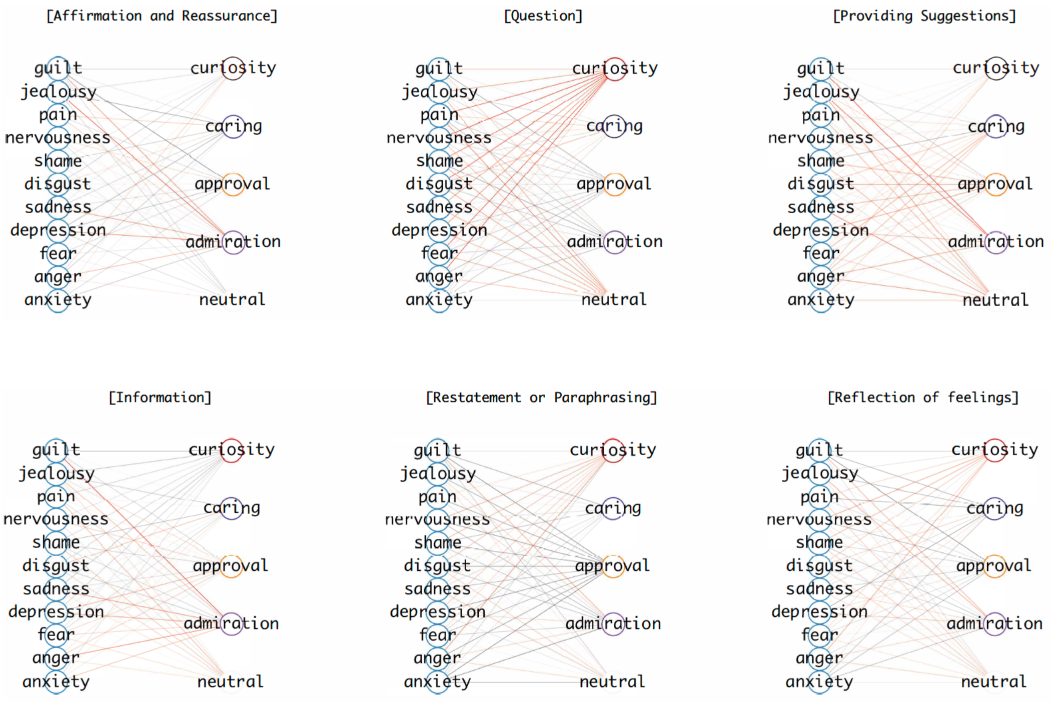

Figure 3: Transition graph in different strategies. Nodes on the right side denote the top-5 seeker emotions, while left-sided nodes denote response verbal emotions. The color depth of each edge indicates the transition probability between a seeker emotion and a response emotion. A red edge denotes a more activated alignment (larger weight) when shifting from "self-disclosure" to each strategy, while a grey edge denotes an inhibited alignment in this shift. The color strength denotes the degree of activation or inhibition.

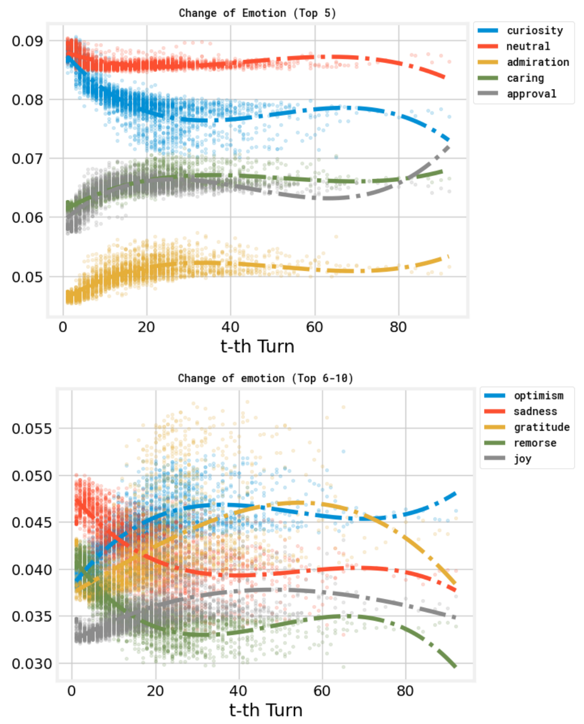

Figure 4: The change of system emotion in the course of emotional support conversation

In the same vein, we notice that our model attentively reacts and shares the user's negative feelings in the second and third examples by uttering “it's hard to be perfect" and “I can imagine ... you are feeling.. stressed". However, the emotional sufferings expressed by the seekers are not carefully appraised or reflected in the two baseline responses.

6. Conclusion

This paper introduces Emstremo, an advanced emotional support system that incorporates emotion-strategy integrated selection. Emstremo effectively captures the delicate emotional transition between seekers and supporters. Our experimental results, evaluated both automatically and by humans, highlight the advantages and merit of Emstremo in providing more empathetic and caring responses. The visualization analysis further suggests the interpretability of our approach in reflecting the intricate sensitive control of emotions in emotional communications. In the future, we will focus on the prediction and control of a wider range of affective features, such as emotional intensity and relational aboutness.

Limitations

While our approach has achieved remarkable advancements compared to the state-of-the-art systems, there still remains a noticeable gap between the generated responses and human production. This is particularly evident in the limited personalization of the automatic responses. Specifically, the automatic responses often remain general and safe, without delving deeper into the seeker's personality, background, and experiences. Additionally, our model frequently produces responses with low answerability, typically when reflecting on the given information in the dialogue without initiating new perspectives to foster ongoing interaction. Future research should consider addressing the issue of answerability to enhance the effectiveness of compute-mediated emotional support.

Looking at the bigger picture, emotional support in conversation is also governed by a couple of demographic and societal factors, typically cultural and social norms. Due to the limitations of data, it is infeasible to incorporate bona fide social-demographic features into our system. We expect future work to enrich the current dataset with reliable social-demographic annotations under sufficient ethical considerations.

Ethical Considerations

The **ESConv** dataset used in this paper is a publically available benchmark that excludes any sensitive or personal information as well as unethical content. Our work aims to provide a potentially rewarding approach to improve the performance of neural and Transformer-based chatbots in emotional support conversations. In this way, the model trained in this paper will not be applied as a replacement for ESC professionals or psychology counseling experts. Furthermore, we prioritize the anonymity and informed consent of participants in our human evaluation.

Acknowledgements

This research was supported by the Large Equipment Fund (project code: 1-BC7N), CBS fund for "Empathetic Dialog Systems" (1-W16H), and “Conversational Brains: A Multidisciplinary Approach” (H-ZG9X), funded by the Hong Kong Polytechnic University. We would also like to thank the four anonymous reviewers for their feedback and suggestions.

Bibliographical References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin

Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge.

Brant R Burleson. 2008. What counts as effective emotional support. *Studies in applied interpersonal communication*, pages 207–227.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. *arXiv preprint arXiv:2306.04657*.

Yixin Chen and Yang Xu. 2021. Exploring the effect of social support and empathy on user engagement in online mental health communities. *International Journal of Environmental Research and Public Health*, 18(13):6855.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. *arXiv preprint arXiv:2210.04242*.

Sheldon Cohen and Thomas A Wills. 1985. Stress, social support, and the buffering hypothesis. *Psychological bulletin*, 98(2):310.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *arXiv preprint arXiv:2305.10172*.

Beverley Fehr. 2004. Intimacy expectations in same-sex friendships: a prototype interaction-pattern model. *Journal of personality and social psychology*, 86(2):265.

Sudeep Gandhe and David Traum. 2008. Evaluation understudy for dialogue coherence models. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 172–181.

Daena J Goldsmith and Kristine Fitch. 1997. The normative context of advice as social support. *Human communication research*, 23(4):454–476.

Amanda J Holmstrom, Brant R Burleson, and Susanne M Jones. 2005. Some consequences for helpers who deliver “cold comfort”: Why it’s worse for women than men to be inept when providing emotional support. *Sex Roles*, 53:153–172.

Bahar Irfan, Anika Narayanan, and James Kennedy. 2020. Dynamic emotional language adaptation in multiparty interactions with agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiazhao Li, Corey Lester, Xinyan Zhao, Yuting Ding, Yun Jiang, and VG Vydiswaran. 2022a. Pharmmt: a neural machine translation approach to simplify prescription directions. *arXiv preprint arXiv:2204.03830*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2019. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022b. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10993–11001.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Zhiqiang Ma, Wenchao Jia, Yutong Zhou, Biqi Xu, Zhiqiang Liu, and Zhuoyi Wu. 2024. Personality enhanced emotion generation modeling for dialogue systems. *Cognitive Computation*, 16(1):293–304.

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. *arXiv preprint arXiv:2010.01454*.

Jingbo Meng and Yue Dai. 2021. Emotional support from ai chatbots: Should a supportive partner self-disclose or not? *Journal of Computer-Mediated Communication*, 26(4):207–222.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. Fado: Feedback-aware double controlling network for emotional support conversation. *Knowledge-Based Systems*, 264:110340.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Towards sentiment and emotion aided multi-modal speech act classification in twitter. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5727–5737.

Andrea Scarantino. 2018. Emotional expressions as speech act analogs. *Philosophy of Science*, 85(5):1038–1053.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

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*.

Erin L Spottswood, Joseph B Walther, Amanda J Holmstrom, and Nicole B Ellison. 2013. Person-centered emotional support and gender attributions in computer-mediated communication. *Human Communication Research*, 39(3):295–316.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: a mixed strategy-aware model integrating comet for emotional support conversation. *arXiv preprint arXiv:2203.13560*.

Bert N Uchino, John T Cacioppo, and Janice K Kiecolt-Glaser. 1996. The relationship between social support and physiological processes: a review with emphasis on underlying mechanisms and implications for health. *Psychological bulletin*, 119(3):488.

Lanrui Wang, Jiangnan Li, Zheng Lin, Fandong Meng, Chenxu Yang, Weiping Wang, and Jie Zhou. 2022. Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. *arXiv preprint arXiv:2210.11715*.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. Transesc: Smoothing emotional support conversation via turn-level state transition. *arXiv preprint arXiv:2305.03296*.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *arXiv preprint arXiv:2307.07994*.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2022. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. *arXiv preprint arXiv:2208.08845*.

Language Resource References

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine

Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. 35(7):6384–6392.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

A. Appendices

A.1. Statistics of ESConv Dataset

In Table 4 and Table 5 are presented the statistics of ESConv Dataset using the official division (Liu et al., 2021) and the MISC division (Tu et al., 2022). In Figure 5 is provided the distribution of strategy labels in the ESConv Dataset.

Category	Official Division		
	Train	Dev	Test
Number of Supporter Utternaces	12759	2722	2895
Avg. words per utterance	18.72	18.83	17.63
Avg. turns per dialogue	23.28	22.91	24.37
Avg. words per dialogue	548.09	548.32	546.16

Table 4: Statistics of ESConv Dataset by the Oficial Division

Category	MISC Division		
	Train	Dev	Test
Number of Supporter Utternaces	14116	1763	1763
Avg. words per utterance	18.16	18.01	18.01
Avg. turns per dialogue	8.61	8.58	8.48
Avg. words per dialogue	156.29	154.58	152.79

Table 5: Statistics of ESConv Dataset by the MISC Division

A.2. Implementation Details

A.2.1. Emotion Classifier

To reach the emotion distribution of ground truth response, we utilize an “off-the-shelf" emotion classifier `SamLowe/roberta-base-go_emotions`, fine-tuned on Go-Emotions Dataset (Demszky et al., 2020) using `FacebookAI/roberta-base`. To use the 28-dimensioned distribution to quantify the emotion state of ground-truth responses. The label set is available in Demszky et al. (2020). The accuracy for each emotion reaches above 0.782 on the test split, demonstrating the model’s capacity to provide emotional representations for ground-truth responses.

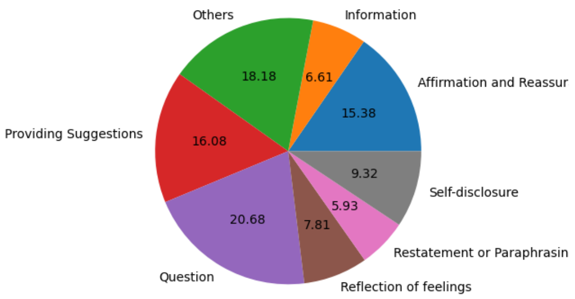

Figure 5: Distribution of Strategy in ESConv

A.2.2. Details of Commensen Knowledge Acquisition

We infer knowledge in nine relations and select the most probable nodes (20 nodes for the situation description and 30 nodes for the user’s last utterance). As the output of the prestrained COMET-ATOMIC model, each knowledge node is formulated as a token sequence. We represent each single node using the encoder block in our model. We prepend to each node the special token denoting the relation leading to the current node. We contact all the “prepended" token sequences and extract the special token’s last hidden state as a knowledge representation. The representations of the 20 or 30 knowledge are stacked (on the dimension of timestep) to form the knowledge-wise hidden state of the situation or user’s final utterance.