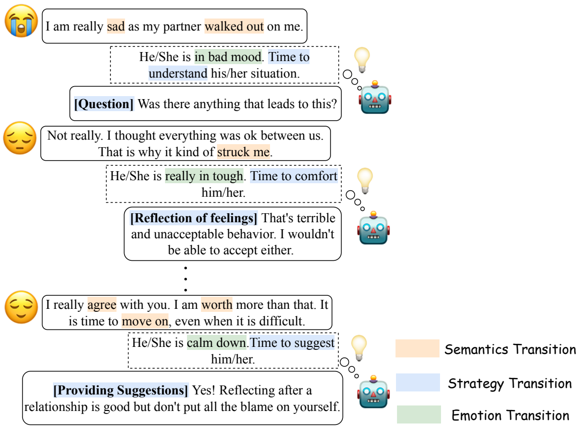
TransESC: Smoothing Emotional Support Conversation via Turn-Level State Transition

Weixiang Zhao, Yanyan Zhao[*], Shilong Wang, Bing Qin
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{wxzhao, yyzhao}@ir.hit.edu.cn

Abstract

Emotion Support Conversation (ESC) is an emerging and challenging task with the goal of reducing the emotional distress of people. Previous attempts fail to maintain smooth transitions between utterances in ESC because they ignore to grasp the fine-grained transition information at each dialogue turn. To solve this problem, we propose to take into account turn-level state **Trans**itions of **ESC** (**TransESC**) from three perspectives, including semantics transition, strategy transition and emotion transition, to drive the conversation in a smooth and natural way. Specifically, we construct the state transition graph with a two-step way, named transit-then-interact, to grasp such three types of turn-level transition information. Finally, they are injected into the transition-aware decoder to generate more engaging responses. Both automatic and human evaluations on the benchmark dataset demonstrate the superiority of TransESC to generate more smooth and effective supportive responses. Our source code is available at https://github.com/circle-hit/TransESC.



I am really sad as my partner walked out on me.

He/She is in bad mood. Time to understand his/her situation.

[Question] Was there anything that leads to this?

Not really. I thought everything was ok between us. That is why it kind of struck me.

He/She is really in tough. Time to comfort him/her.

[Reflection of feelings] That's terrible and unacceptable behavior. I wouldn't be able to accept either.

I really agree with you. I am worth more than that. It is time to move on, even when it is difficult.

He/She is calm down.Time to suggest him/her.

[Providing Suggestions] Yes! Reflecting after a relationship is good but don't put all the blame on yourself.

Figure 1: An example for the turn-level state transitions during an emotional support conversation from the ESConv (Liu et al., 2021) dataset.

1 Introduction

Emotional Support Conversation (ESC) is a goal-directed task which aims at reducing individuals' emotional distress and bringing about modifications in the psychological states of them. It is a desirable and critical capacity that an engaging chatbot is expected to have and has potential applications in several areas such as mental health support, customer service platform, etc.

Different from the emotional (Zhou et al., 2018) and empathetic (Rashkin et al., 2019) conversation, ESC is always of long turns, which requires skillful conversation procedures and support strategies to achieve the goal. For example, as shown in Figure 1, the supporter should firstly explore the situation to identify the problems faced by the seeker, and then try to comfort him. In the end, helpful suggestions are provided to help the seeker get rid of the tough. Intuitively, for such a complex and challenging task, a question is left: *how to maintain smooth transitions between utterances from different procedures and drive the conversation in a natural way?* Previous works (Liu et al., 2021; Peng et al., 2022; Tu et al., 2022) fail to deal with this issue because they treat the dialogue history as a long sequence, ignoring to grasp the fine-grained transition information at each dialogue turn. We argue that considering such turn-level transition information plays the crucial role in achieving effective ESC, navigating the conversation towards the expected goal to reduce the seeker's distress in a smooth way. To achieve this, we model the transition information in ESC from three perspectives and refer to each one of them as a state.

First, it is a common phenomena that, even focusing on the same topic, the help seeker may tell different aspects or meanings as the conversation goes. We refer to it as **semantics transition** and take the example in Figure 1. To begin with, the

[*] Corresponding author

help seeker feels sad to break up with the partner and does not know the reason (e.g. *sad*, *walked out*, *struck me*). After receiving the warm and skillful emotional support from the supporter, he is relieved and encouraged to move forward (e.g. *agree*, *worth*, *move on*). Thus, to fully comprehend the dialogue content with the goal of achieving effective emotional support, it is crucial to grasp such fine-grained semantic changes at each dialogue turn.

Second, the timing to adopt proper support strategies constitutes another important aspect to achieve effective emotional support. In Figure 1, the supporter attempts to understand the seeker's problem via a *Question* and comfort him by *Reflection of feelings*. And the emotional support ends with the strategy *Providing Suggestion* to help the seeker get through the tough. Such flexible combination and dependencies of different strategies forms the **strategy transition** in ESC, driving the conversation in the more natural and smooth way to solve the dilemma faced by the seeker.

Finally, it is also of vital importance to track the emotional state of the seeker as conversation develops. The seeker in Figure 1 comes with a *bad mood* and suffers from the *tough* that his partner chooses to leave. As the ESC goes, his emotional state is changed and becomes *calm down* to move on. Grasping such **emotion transition** can provide the supporter clear signals to apply proper strategies and offer immediate feedbacks to be aware of the effectiveness of the emotional support..

In this paper, in order to maintain smooth transitions between utterances in ESC and drive the conversation in a natural way, we propose to take into account turn-level state **Trans**itions of **ESC** (**TransESC**), including semantics transition, strategy transition and emotion transition. To be more specific, we construct the state transition graph for the process of emotional support. Each node consists of three types of states, representing semantics state, strategy state and emotion state of the seeker or the supporter at each dialogue turn. And seven types of edges form the path for information flow. Then we devise a two-step way, called transit-then-interact, to explicitly perform state transitions and update each node representation. During this process, ESC is smoothed through turn-level supervision signal that keywords of each utterance, adopted strategies by the support and immediate emotional states of the seeker are predicted by the corresponding state representations at each turn. Finally, we inject the obtained three transition information into the decoder to generate more engaging and effective supportive response.

The main contributions of this work are summarized as follows:

- We propose to smooth emotional support conversation via turn-level state transitions, including semantics transition, strategy transition and emotion transition.

- We devise a novel model TransESC to explicitly transit, interact and inject the state transition information into the process of emotional support generation.

- Results of extensive experiments on the benchmark dataset demonstrate the effectiveness of TransESC to select the exact strategy and generate more natural and smooth responses.

2 Related Works

2.1 Emotional Support Conversation

Liu et al. (2021) propose the task of emotional support conversation and release the benchmark dataset ESConv. They append the support strategy as a special token into the beginning of each supportive response and the following generation process is conditioned on the predicted strategy token. Peng et al. (2022) propose a hierarchical graph network to utilize both the global emotion cause and the local user intention. Instead of using the single strategy to generate responses, Tu et al. (2022) incorporate commonsense knowledge and mixed response strategy into emotional support conversation. More recently, Cheng et al. (2022) propose look-ahead strategy planning to select strategies that can lead to the best long-term effects and Peng et al. (2023) attempt to select an appropriate strategy with the feedback of the seeker. However, all existing methods treat the dialogue history as a lengthy sequence and ignore the turn-level transition information that plays critical roles in driving the emotional support conversation in a more smooth and natural way.

2.2 Emotional & Empathetic Conversation

Endowing emotion and empathy to the dialogue systems has gained more and more attentions recently. To achieve the former goal, both generation-based methods (Zhou et al., 2018; Zhou and Wang, 2018; Shen and Feng, 2020) and retrieval-based

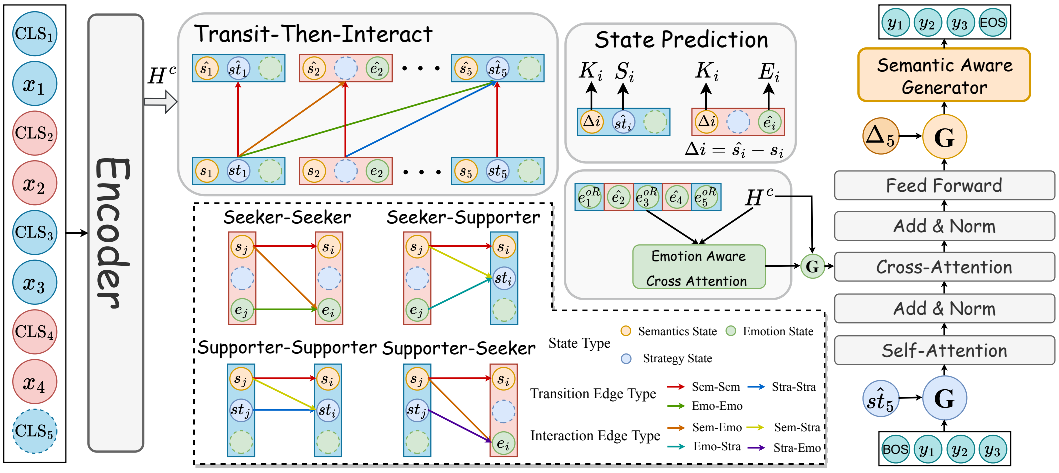


Figure 2: The overall architecture of our proposed TransESC model, which mainly consists of three modules: Context Encoder, Turn-Level State Transition Module and Transition-Aware Decoder.

(Qiu et al., 2020; Lu et al., 2021) methods attempt to incorporate emotion into dialogue generation. However, it merely meets the basic quality of dialog systems. And to generate empathetic response, previous works incorporate affection (Alam et al., 2018; Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Li et al., 2020, 2022), cognition (Sabour et al., 2022; Zhao et al., 2022) or persona (Zhong et al., 2020) aspects of empathy. Intuitively, expressing empathy is only one of the necessary steps to achieve effective emotional support. By contrast, emotional support is a more high-level ability that dialogue systems are expected to have.

3 Preliminaries

3.1 ESConv Dataset

Our research is carried out on the **E**motional **S**upport **Conv**ersation dataset, ESCONV (Liu et al., 2021). In each conversation, the seeker with a bad emotional state seeks help to go through the tough. And the supporter is supposed to identify the problem that the seeker is facing, console the seeker, and then provide some suggestions to help the seeker to overcome their problems. The support strategies adopted by the supporter are annotated in the dataset and there are eight types of strategies (e.g., *question*, *reflection of feelings* and *providing suggestions*). However, ESCONV dataset does not contain keyword sets of each utterance and emotion labels [1] for the seeker's turn, we leverage external tools to automatically annotate them. More details about annotation are provided in Appendix A.

3.2 Task Definition

Formally, let $D = [X_1, X_2, \cdots, X_N]$ denotes a dialogue history with N utterances between the seeker and the supporter, where the i-th utterance $X_i = [w_1^i, w_2^i \cdots, w_m^i]$ is a sequence of m words. And each utterance is provided with the extracted set of top k keywords $K_i = [k_1^i, k_2^i \cdots, k_k^i]$. Besides, the adopted support strategy S_i of the supporter and the emotional state label E_i of the seeker are also available for the turn-level supervision. The goal is to generate the next utterance Y from the stand of the supporter that is coherent to the dialogue history D and supportive to reduce the seeker's distress.

4 Methodology

The overall architecture of our proposed TransESC is shown in Figure 2. The dialogue representations are first obtained through context encoder. Then we grasp and propagate the fine-grained transition information, including semantics transition, strategy transition and emotion transition, in the Turn-Level State Transition Module. Finally, to generate more natural and smooth emotional support responses, such transition information is clearly injected into the Transition-Aware Decoder.

4.1 Context Encoder

We adopt Transformer encoder (Vaswani et al., 2017) to obtain the contextual representations of the dialogue history. Following previous works (Tu et al., 2022), the dialogue is flattened into a word sequence. Then we append the special token [CLS] to the beginning of each utterance and another one for

[1] We use 6 emotion categories: joy, anger, sadness, fear, disgust, and neutral.

the upcoming response. And the context encoder produces the contextual embeddings $H^c \in \mathbb{R}^{N \times d_h}$.

4.2 Turn-Level State Transition

In this section, we propose to grasp the turn-level transition information, including semantics transition, strategy transition and emotion transition, to explicitly smooth the emotional support and drive the conversation in a natural way. Specifically, we construct the state transition graph, with three types of state for each node and seven types of edges, to propagate and update the transition information. And all the three states are supervised at each dialogue turn to predict the keyword set of each utterance, the adopted strategy of the supporter and the emotional state of the seeker.

State Transition Graph. We construct the state transition graph to grasp and propagate transition information at each dialogue turn. To alleviate the impact of lengthy and redundant dialogue history, we perform the state transition within a fixed window size w. Specifically, we regard the current turn of supporter's response u_e as the end and the w-th latest utterance u_s spoken by the supporter as the start. All the utterances between u_s and u_e constitute the transition window.

Nodes: There are three types of states in total, making up each node in the transition graph. Since the adopted strategy and the emotional state are specified for the supporter and the seeker respectively, for the nodes from the supporter's turn, they include the semantics state and the strategy state, while the semantics state and the emotion state constitute the nodes for the seeker's turn.

Edges: We build edges to connect each node with all previous ones. Since there are two roles in ESC, it leads to four types of connection ways (e.g. Seeker-Seeker) between any two nodes. And seven types of edge are divided into two groups, the transition edges $\mathcal{T}$ and the interaction edges $\mathcal{I}$. For the former ones, they function to transit previous influences and grasp dependencies between states of the same type (e.g. Strategy-Strategy), while the later ones are devised to perform the interaction between different state types (e.g. Strategy-Emotion). The idea behind the interaction types is that decisions of the supporter to choose a certain strategy should focus on what the seeker said and are largely determined by emotional states of him/her. Also, what the supporter expressed and the adopted strategy could directly have impact on the emotional state of the seeker, leading the seeker into the better mood.

Graph Initialization. Here we introduce the way to initialize three states for each node.

For the **semantics state** and the **strategy state** of each node, they are both initialized by the corresponding $[\mathrm{CLS_i}]$ token of each utterance.

And for the **emotion state**, in addition to initialized by the $[\mathrm{CLS_i}]$ token, we also leverage commonsense knowledge from the external knowledge base ATOMIC (Sap et al., 2019) to imply the emotional knowledge of the seeker at each dialogue turn. Concretely, the generative commonsense transformer model COMET (Bosselut et al., 2019) is adopted to obtain the knowledge. We select relation type *xReact* to manifest the emotional feelings of the seeker. Then the hidden state representations from the last layer of COMET are obtained as the emotional knowledge csk_i. The final representation of the emotion state is the sum of $[\mathrm{CLS_i}]$ and csk_i. Please refer to the Appendix B for the detailed implementation of COMET and definitions of the knowledge relation types in ATOMIC.

Transit-Then-Interact. In order to explicitly grasp the turn-level transition information of the three states, we devise the two-step way Transit-Then-Interact (TTI) to propagate and update state representations of each node. Specifically, inspired by Li et al. (2021a), the relation-enhanced multi-head attention (MHA) (Vaswani et al., 2017) is applied to update node representations from the information of the connected neighbourhoods. The formulation of vanilla MHA could be written as:

$$\hat{v}_i = \underset{j \in \mathcal{N}}{\mathrm{MHA}}(q_i, k_j, v_j), \tag{1}$$

where $\mathrm{MHA}(Q, K, V)$ follows the implementation of multi-head attention (Vaswani et al., 2017)

And the key of relation-enhanced multi-head attention (R-MHA) is that we incorporate the embeddings of edge types into the query and the key. Thus, the two-step Transit-Then-Interact process operated on semantics states could be written as:

$$s'_i = \underset{e_{ij} \in \mathcal{T}}{\text{R-MHA}}(s_i + r_{ij}, s_j + r_{ij}, s_j), \tag{2}$$

$$s''_i = \underset{e_{ij} \in \mathcal{I}}{\text{R-MHA}}(s'_i + r_{ij}, s'_j + r_{ij}, s'_j), \tag{3}$$

where e_{ij} is the edge type between the semantics states at i-th turn and that of j-th turn. $\mathcal{T}$ and $\mathcal{I}$ are the transition edge types and the interaction edge types, respectively. r_{ij} is the embedding of e_{ij}.

Then we dynamically fuse the results of transition s_i' and interaction s_i'' to obtain the updated semantics state $\hat{s}_i$:

$$\begin{aligned} \hat{s}_i &= g^{tti} \odot s_i' + (1 - g^{tti}) \odot s_i'' \\ g^{tti} &= \sigma([s_i'; s_i'']W^{tti} + b^{tti}) \end{aligned} \tag{4}$$

where $W^{tti} \in \mathbb{R}^{2d_h \times d_h}$ and $b^{tti} \in \mathbb{R}^{d_h}$ are trainable parameters.

Similarly, the ways to obtain the updated strategy state $\hat{st}_i$ and emotion state $\hat{e}_i$ are identical to that of the above semantics state $\hat{s}_i$.

4.3 State Prediction

We utilize the turn-level annotation to supervise the transition information, driving the emotional support conversation in a smooth and natural way.

Semantic Keyword Prediction. In order to measure the semantics transition more concretely, inspired by Li et al. (2021b), we calculate the difference $\Delta_i = \hat{s}_i - s_i$ between the semantics state before and after the operation TTI. Then we devise a bag-of-words loss to force Δ_i to predict the semantics keyword set $K_i = [k_1^i, k_2^i \cdots, k_k^i]$ of the corresponding utterance.

$$\begin{aligned} \mathcal{L}_{SEM} &= -\sum_{i=1}^{N}\sum_{j=1}^{k} \log p(k_j^i|\Delta_i) \\ &= -\sum_{i=1}^{N}\sum_{j=1}^{k} \log f_{k_j^i} \end{aligned} \tag{5}$$

where $f_{k_j^i}$ denotes the estimated probability of the j-th keyword k_j^i in the utterance u_i. The function f serves to predict the keyword set of the utterance u_i in a non-autoregressive way:

$$f = \text{softmax}(W^{sem}\Delta_i + b^{sem}) \tag{6}$$

where $W^{sem} \in \mathbb{R}^{d_h \times |V|}$, $b^{sem} \in \mathbb{R}^{|V|}$ and V refers to the vocabulary size.

Supporter Strategy Prediction. After the TTI module, we attempt to explicitly model the dependencies among the adopted supportive strategy during the ESC. Then we utilize the strategy label S_i to specify the strategy state at each dialogue turn.

$$\hat{y}_{str} = \text{softmax}(W^{str}\hat{st}_i + b^{str}) \tag{7}$$

where $\hat{y}_{str} \in \mathbb{R}^{n_s}$, $W^{str} \in \mathbb{R}^{d_h \times n_s}$ and $b^{sem} \in \mathbb{R}^{n_s}$. n_s is the number of total available strategy.

Cross entropy loss is utilized and the loss function is defined as:

$$\mathcal{L}_{STR} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_s} \hat{y}_{str,i}^j \cdot log(y_{str,i}^j) \tag{8}$$

where $y_{str,i}^j$ stands for the ground-truth strategy label of the utterance i from the supporter.

Seeker Emotion Prediction. Similarly, the emotion states e_i of each seeker's dialogue turn are also fed into another linear transformation layer:

$$\hat{y}_{emo} = \text{softmax}(W^{emo}\hat{e}_i + b^{emo}) \tag{9}$$

where $\hat{y}_{emo} \in \mathbb{R}^{n_e}$, $W^{emo} \in \mathbb{R}^{d_h \times n_e}$ and $b^{emo} \in \mathbb{R}^{n_e}$. n_e is the number of total available emotion.

Cross entropy loss is also utilized for training:

$$\mathcal{L}_{EMO} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_e} \hat{y}_{emo,i}^j \cdot log(y_{emo,i}^j) \tag{10}$$

where $y_{emo,i}^j$ is the ground-truth emotion label of the utterance i from the seeker.

4.4 Transition-Aware Decoder

Finally, based on the vanilla Transformer decoder (Vaswani et al., 2017), we devise the transition aware decoder to inject the turn-level transition information into the process of response generation.

To make the generation process grounded on the selected strategy, we dynamically fuse the last strategy state $\hat{st}$ (the adopted strategy for the upcoming response) with the embeddings of the utterance sequence as the input of the decoder:

$$\begin{aligned} \hat{E}_i &= g^{str} \odot E_i + (1 - g^{str}) \odot \hat{st} \\ g^{str} &= \sigma([E_i; \hat{st}]W^1 + b^1) \end{aligned} \tag{11}$$

where $W^1 \in \mathbb{R}^{2d_h \times d_h}$ and $b^1 \in \mathbb{R}^{d_h}$ are trainable parameters and E_i is the i-th embedding token of the response.

And for the emotion transition information, we dynamically combine it with the output of the context encoder H^c to explicitly incorporate the emotional states of the seeker. Specifically, the emotion states e_i of the seeker and commonsense knowledge e_i^{oR} of the supporter, which is generated by the COMET model under the relation type *oReact* to imply what the emotional effect would exert on the seeker after the i-th utterance of the supporter, constitutes the emotional state sequence H^{emo}.

$$\begin{aligned} \hat{H} &= g^{emo} \odot H^c + (1 - g^{emo}) \odot \hat{H}^{emo} \\ \hat{H}^{emo} &= \text{Cross-Att}(H^c, H^{emo}) \\ g^{emo} &= \sigma([H^c; \hat{H}^{emo}]W^2 + b^2) \end{aligned} \tag{12}$$

where $W^2 \in \mathbb{R}^{2d_h \times d_h}$ and $b^2 \in \mathbb{R}^{d_h}$ are trainable parameters.

Thus, for the target response $Y = [y_1, y_2, \cdots, y_M]$, to generate the t-th token y_t, the hidden representation of it from the decoder can be obtained:

$$h_t = \text{Decoder}(\hat{E}_{y<t}, \hat{H}) \tag{13}$$

In the end, we dynamically inject semantics transition information via the fusion of the last semantics difference representation Δ_i (latent semantic information for the upcoming utterance) and the hidden representation h_t of the t-th token:

$$\begin{aligned} \hat{h} &= g^{sem} \odot h_t + (1 - g^{sem}) \odot \Delta_i \\ g^{sem} &= \sigma([h_t; \Delta_i]W^{sem} + b^{sem}) \end{aligned} \tag{14}$$

where $W^3 \in \mathbb{R}^{2d_h \times d_h}$ and $b^3 \in \mathbb{R}^{d_h}$ are trainable parameters.

The distribution over the vocabulary for the t-th token can be obtained by a softmax layer:

$$P(y_t \mid y_{<t}, D) = \text{softmax}(W\hat{h} + b) \tag{15}$$

where D is the input dialogue history.

We utilise the standard negative log-likelihood as the response generation loss function:

$$L_{gen} = -\sum_{t=1}^{M} \log P\left(y_t \mid D, y_{<t}\right). \tag{16}$$

A multi-task learning framework is adopted to jointly minimize the response generation loss, the semantic keyword, strategy and emotion loss.

$$\mathcal{L} = \gamma_1 \mathcal{L}_{GEN} + \gamma_2 \mathcal{L}_{SEM} + \gamma_3 \mathcal{L}_{STR} + \gamma_4 \mathcal{L}_{EMO} \tag{17}$$

where γ_1, γ_2, γ_3 and γ_4 are hyper-parameters.

5 Experiments

5.1 Baselines

We compare our proposed TransESC with the following competitive baselines. They are four empathetic response generators: **Transformer** (Vaswani et al., 2017), **Multi-Task Transformer (Multi-TRS)** (Rashkin et al., 2019), **MoEL** (Lin et al., 2019) and **MIME** (Majumder et al., 2020); and two state-of-the-art models on ESC task: **BlenderBot-Joint** (Liu et al., 2021), **GLHG** (Peng et al., 2022) and **MISC** (Tu et al., 2022). More details of them are described in Appendix C.

5.2 Implementation Details

To be comparable with baselines, we implement our model based on BlenderBot-small (Roller et al., 2021) with the size of 90M parameters. The window size w of turn-level transition is 2. The hidden dimension d_h is set to 300 and the number of attention heads in relation enhanced multi-head attention and emotion aware attention graph are 16 and 4. Loss weights γ_1, γ_2, γ_3 and γ_4 are set to 1, 0.2, 1 and 1, respectively. AdamW (Loshchilov and Hutter, 2017) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ is used for training. We vary the learning rate during the training process with the initial learning rate of 2e-5 and use a linear warmup with 120 warmup steps. And the training process is performed on one single NVIDIA Tesla A100 GPU with a mini-batch size of 20. For inference, following Tu et al. (2022), we also adopt the decoding algorithms of Top-p and Top-k sampling with p=0.3, k=30, temperature τ=0.7 and the repetition penalty 1.03.

5.3 Evaluation Metrics

Automatic Evaluation. We apply four kinds of automatic metrics for evaluation: (1) Perplexity (**PPL**) measures the general quality of the generated responses; (2) BLEU-2 (**B-2**), BLEU-4 (**B-4**) (Papineni et al., 2002) and ROUGE-L (**R-L**) (Lin, 2004) evaluate the lexical and semantic aspects of the generated responses; (3) Distinct-n (**Dist**-n) (Li et al., 2016) evaluates the diversity of the generated responses by measuring the ratio of unique n-grams; (4) Accuracy (**Acc**) of the strategy prediction is utilised to evaluate the model capability to choose the supportive strategy.

Human Evaluation. Following Liu et al. (2021), we recruit three professional annotators to interact with the models for human evaluation. Specifically, 100 dialogues from the test set of ESConv are randomly sampled. Then we ask the annotators to act as seekers under these dialogue scenarios and chat with the models. Given TransESC and a compared model, the annotators are required to choose which one performs better (or tie) following five

Model	Acc	PPL	D-1	D-2	B-1	B-2	B-3	B-4	R-L
Transformer	-	89.61	1.29	6.91	-	6.53	-	1.37	15.17
Multi-TRS	-	89.52	1.28	7.12	-	6.58	-	1.47	14.75
MoEL	-	133.13	2.33	15.26	-	5.93	-	1.22	14.65
MIME	-	47.51	2.11	10.94	-	5.23	-	1.17	14.74
BlenderBot-Joint	17.69	17.39	2.96	17.87	18.78	7.02	3.20	1.63	14.92
GLHG	-	**15.67**	3.50	**21.61**	**19.66**	7.57	3.74	2.13	16.37
MISC	31.67	16.27	4.62	20.17	16.31	6.57	3.26	1.83	17.24
TransESC (Ours)	**34.71**	15.85	**4.73**	20.48	17.92	**7.64**	**4.01**	**2.43**	**17.51**

Table 1: Comparison of our model against state-of-the-art baselines in terms of the automatic evaluation. The best results among all models are highlighted in bold.

TransESC vs.	BlenderBot-Joint			MISC		
	Win	Lose	Tie	Win	Lose	Tie
Fluency	**54.7**‡	18.0	27.3	**65.7**‡	10.7	23.7
Identification	**37.3**‡	16.0	46.7	**32.0**	19.3	48.7
Empathy	**39.3**‡	7.0	53.7	**48.0**‡	5.7	46.3
Suggestion	**37.0**	27.7	35.3	**46.7**†	17.3	36.0
Overall	**51.7**‡	26.0	22.3	**64.0**‡	17.7	18.3

Table 2: The results of the human interaction evaluation (%). TransESC performs better than all other models (sign test, ‡ / † represent p-value < 0.05 / 0.1).

Model	Dist-1	B-2	B-4	R-L
TransESC	4.73	**7.64**	**2.43**	**17.51**
w/o Sem. Trans	4.55	7.04	2.13	17.37
w/o Stra. Trans	4.29	6.68	2.01	17.15
w/o Emo. Trans	**4.82**	7.14	2.22	17.45
w/o T-L. Trans	4.19	6.35	1.94	16.88

Table 3: Results of ablation study. Sem./Stra./Emo./T-L. Trans refer to the semantics/strategy/emotion/all three types of turn-level transition, respectively.

aspects: (1) **Fluency**: which model generates more coherent and smooth responses; (2) **Identification**: which model explores the seeker's problems more effectively; (3) **Empathy**: which model is more empathetic to understanding the seeker's feelings and situations; (4) **Suggestion**: which model offers more helpful suggestions; (5) **Overall**: which model provides more effective emotional support.

6 Results and Analysis

6.1 Overall Results

Automatic Evaluation. As shown in Table 2, TransESC achieves the new state-of-the-art automatic evaluation results. Benefiting from the grasp of three types of transition information in ESC, TransESC is capable of generating more natural and smooth emotional support responses in terms of almost all the metrics compared to the baselines. Compared with the empathetic response generators, the significant performance gain of TransESC demonstrates that eliciting empathy is only one of the critical procedures of ESC, while identifying the problems faced by the seeker and offering helpful suggestions also constitute the important aspects in ESC. Moreover, although the process of strategy prediction is also explored in BlenderBot-Joint and MISC, the prominent performance on strategy selection of TransESC can be ascribed to the explicit turn-level strategy transition modeling, which sufficiently capture the dependencies of different strategies adopted at each supporter's turn. As shown in Figure 3, TransESC also outperforms baselines in terms of all the top-n accuracy.

Human Evaluation. For the evaluation setting, it is worth to mention that MISC takes the pre-conversation "situation" of the seeker as the input, which is not rational because the supporter can only comprehend what the seeker is facing as conversation goes. Thus, for the fair comparison, we do not input the "situation" for all three models. As shown in Table 2, TransESC outperforms them in terms of all evaluation aspects. Specifically, it generates more fluent and smooth responses in terms of higher Fluency score, which verifies the benefits of incorporating turn-level transition information to maintain smooth transition between utterances. Also, although all three models may be comparable to identify problems of the seeker, TransESC could elicit more empathetic responses to comfort the seeker and then offer more helpful suggestions.

6.2 Ablation Study

To explore the impact of three types of transition information, we remove the corresponding state representation with edges in the transition graph, the

Situation	There is no hope, I am struggling with the pandemic and loneliness
Context	**Supporter**: **[Affirmation and Reassurance]** I know that days can be really hard. I think ... **Seeker**: Yeah, I just kind of feel like a failure in life **Seeker**: But I am trying, thanks. **Supporter**: **[Affirmation and Reassurance]** I understand that there are things in your life ...
BlenderBot-Joint	**[Self-disclosure]** I can understand why you are feeling this way. It is very difficult to see people be put down for the things that are bothering you.
MISC	**[Others]** I think you are doing the right thing!
TransESC	**[Providing Suggestions]** I think that you should try to focus on what is important to you. I know it can be hard to do that when you are feeling down but I believe that you can do it!
Ground-Truth	**[Providing Suggestions]** When you feel up to it, do a search for temp agencies near you and hopefully they can give you some leads about a job.

Table 4: Case study of the generated supportive responses by our proposed TransESC and the baselines.

Win. Size	Dist-1	B-2	B-4	R-L
$w = 1$	4.68	7.49	2.27	17.25
$w = 2$	**4.73**	**7.64**	**2.43**	**17.51**
$w = 3$	4.49	6.52	2.26	17.29
$w = 4$	4.39	7.04	2.12	17.29
$w = 5$	4.71	6.98	2.17	17.24

Table 5: Results of our proposed model with different lengths of transition window w.

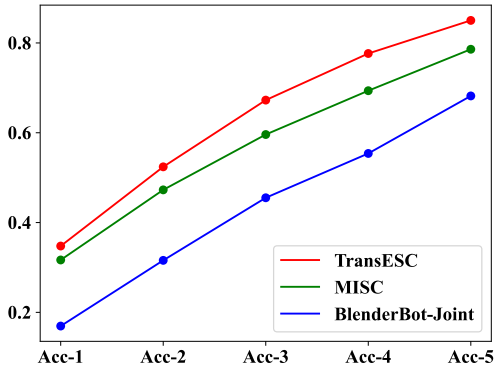


Figure 3: The top-n strategy prediction accuracy of TransESC and two baseline models.

turn-level label prediction and the injection into the decoder. Besides, to explore the effect of turn-level transition process, we also discard it by predicting three states with the whole dialogue history.

As shown in Table 3, the ablation of any types of transition information can lead to a drop in the automatic evaluation results, demonstrating the effectiveness of each one of them. To be more specific, the ablation of the strategy transition (w/o Stra.Trans) causes the most significant performance drop. The reason is that selecting the proper strategy to support the seeker plays the most pivotal role in ESC. And the impact of emotion transition (w/o Emo.Trans) is relatively small. It may be attributed to the noise of annotated emotion labels and the generated emotional knowledge.

Moreover, when we remove the whole process of turn-level state transition, the significant performance drop verifies our contribution that grasping the fine-grained transition information can drive the ESC in a more smooth and natural way.

6.3 Case Study

In Table 4, we show a case with responses generated by TransESC and two baselines. With the emotion transition and strategy transition, after several turns of comforting, TransESC senses the emotion state *joy* of the seeker and it is time to offer helpful suggestions with the correct predicted strategy. And through semantics transition, it grasp the determination of the seeker to suggest him to have a try and encourage him to face the failure. By contrast, MISC and BlenderBot-Joint drive the conversation improperly, leading to the ineffective responses.

6.4 Length of Transition Window

We adjust different lengths of transition window for a deeper analysis of the impact of transition information modeling. Results are shown in Table 5. The model with the transition window length of 2 achieves the best performance. On the one hand, capturing the transition information in the shorter window could not sufficiently comprehend dependencies of utterance transition in the dialogue history. On the other hand, much more redundant transition information may be incorporated by the model with longer transition window, which would weaken the performance of our model.

7 Conclusion and Future Work

In this paper, we propose TransESC to generate emotional support via turn-level state transition information incorporated, including semantics transition, strategy transition and emotion transition. We construct the transition graph with the two-step way, transit-then-interact, to grasp and supervise the transition information at each dialogue turn. Experimental results on both automatic and human evaluation demonstrate the superiority of TransESC to generate more smooth responses.

In the future, we will explore more characteristics in ESC such as persona to generate more natural responses.

8 Limitations

Although our proposed method exhibits great performance to generate more smooth and natural emotional support than baseline models, we argue that the research on this field still has a long way to go. We conclude three aspects that may inspire further exploration. First, the automatically annotated emotion labels may be a little bit coarse and may not accurately manifest the emotional states of the seeker. Second, since various types of commonsense knowledge are not introduced, the current chatbots always generate general and safe responses, failing to provide specific and personalized suggestions to help the seeker get over the dilemma. Finally, current automatic evaluation metrics are still not rational and proper to measure the ability of chabots to provide emotional support. It is desirable to build better evaluation metrics for this.

9 Ethics Statement

The open-source benchmark dataset ESConv (Liu et al., 2021) used in our experiments is well-established and collected by employed crowd-sourced workers, with user privacy protected and no personal information involved. And for our human evaluation, all participants are volunteered and transparently informed of our research intent, with reasonable wages paid.

Moreover, our research only focuses on building emotional support systems in daily conversations, like the one to seek the emotional support from our friends or families. It is worth to mention that we do not claim to construct chatbots that can provide professional psycho-counseling or professional diagnosis. This requires particular caution and further efforts to construct a safer emotional support system, which is capable of detecting users who have tendencies of self-harming or suicide.

Acknowledgements

We thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key RD Program of China via grant 2021YFF0901602 and the National Natural Science Foundation of China (NSFC) via grant 62176078.

References

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

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*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. 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, ACL 2020, Online, July 5-10, 2020*, pages 4040–4054. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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.

In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Junyi Li, Wayne Xin Zhao, Zhicheng Wei, Nicholas Jing Yuan, and Ji-Rong Wen. 2021a. Knowledge-based review generation by coherence enhanced text planning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 183–192. ACM.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4454–4466. International Committee on Computational Linguistics.

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

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021b. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 128–138. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

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 and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Xin Lu, Yijian Tian, Yanyan Zhao, and Bing Qin. 2021. Retrieve, discriminate and rewrite: A simple and effective framework for obtaining affective response in retrieval-based chatbots. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1956–1969. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. 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 2020, Online, November 16-20, 2020*, pages 8968–8979. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

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.

Lisong Qiu, Yingwai Shiu, Pingping Lin, Ruihua Song, Yue Liu, Dongyan Zhao, and Rui Yan. 2020. What if bots feel moods? In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 1161–1170. ACM.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: commonsense-aware empathetic response generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11229–11237. AAAI Press.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Lei Shen and Yang Feng. 2020. CDL: curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 556–566. Association for Computational Linguistics.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 308–319. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2022. Don't lose yourself! empathetic response generation via explicit self-other awareness. *arXiv preprint arXiv:2210.03884*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1128–1137. Association for Computational Linguistics.

Category	Train	Dev	Test
# dialogues	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 6: The statistics of processed ESConv dataset.

A ESCONV Dataset

A.1 Keyword and Emotion Annotation

Since the original ESCONV dataset does not contain keyword sets of each utterance and emotion labels for the seeker's turn, we leverage external tools to annotate them. To obtain the keyword set of each utterance, we use TF-IDF method. The vocabulary and IDF term are learned from the training set of ESCONV. Then for each utterance, we apply TF-IDF to obtain the top k keywords.

For the emotion labels, we fine-tune the BERT model (Devlin et al., 2019) on a fine-grained emotion classification dataset, GoEmotions (Demszky et al., 2020). The the finetuned BERT model achieve an accuracy of 71% on test set, indicating that it is reliable for emotion classification. Then it is used to annotate an emotion label for each utterance from the seeker's turn.

A.2 Dataset Statistics

We carry out the experiments on the dataset ESCONV (Liu et al., 2021) [2]. For pre-processing, following (Tu et al., 2022) we truncate the conversation examples every 10 utterances, and randomly spilt the dataset into train/valid/test set with the ratio of 8:1:1. The statistics is given in Table 6.

A.3 Definitions of Strategies

There are overall 8 types of support strategies that are originally annotated in the ESCONV dataset:

- **Question**: ask for information related to the problem to help the help-seeker articulate the issues that they face.

- **Restatement or Paraphrasing**: a simple, more concise rephrasing of the support-seeker's statements that could help them see their situation more clearly.

- **Reflection of Feelings**: describe the help-seeker's feelings to show the understanding of the situation and empathy.

- **Self-disclosure**: share similar experiences or emotions that the supporter has also experienced to express your empathy.

- **Affirmation and Reassurance**: affirm the help-seeker's ideas, motivations, and strengths to give reassurance and encouragement.

- **Providing Suggestions**: provide suggestions about how to get over the tough and change the current situation.

- **Information**: provide useful information to the help-seeker, for example with data, facts, opinions, resources, or by answering questions.

- **Others**: other support strategies that do not fall into the above categories.

B Commonsense Knowledge Acquisition

B.1 Description of ATOMIC Relations

ATOMIC (Sap et al., 2019) is an atlas of everyday commonsense reasoning and organized through textual descriptions of inferential knowledge, where nine if-then relation types are proposed to distinguish causes vs. effects, agents vs. themes, voluntary vs. involuntary events, and actions vs. mental states. We give the brief definition of each relation.

- **xIntent**: Why does PersonX cause the event?

- **xNeed**: What does PersonX need to do before the event?

- **xAttr**: How would PersonX be described?

- **xEffect**: What effects does the event have on PersonX?

- **xWant**: What would PersonX likely want to do after the event?

- **xReact**: How does PersonX feel after the event?

- **oReact** How does others' feel after the event?

- **oWant** What would others likely want to do after the event?

- **oEffect** What effects does the event have on others?

[2] https://github.com/thu-coai/Emotional-Support-Conversation

B.2 Implementation Details of COMET

The generative commonsense transformer model COMET (Bosselut et al., 2019) is adopted to obtain the knowledge. We select relation types *xReact* to manifest the emotional feelings of the seeker at each dialogue turn. Specifically, we adopt the BART-based (Lewis et al., 2020) variation of COMET, which is trained on the ATOMIC-2020 dataset (Hwang et al., 2021). And given each utterance X_i belonging to the self to form the input format $(X_i, r, \text{[GEN]})$, COMET would generate descriptions of inferential content under the relation r. Then the hidden state representations from the last layer of COMET are obtained as knowledge representation.

C Baselines

- **Transformer** (Vaswani et al., 2017): The vanilla Transformer-based encoder-decoder generation model.

- **Multi-Task Transformer (Multi-TRS)** (Rashkin et al., 2019): A variation of the vanilla Transformer with an auxiliary task to perform emotion perception of the user.

- **MoEL** (Lin et al., 2019): A Transformer-based model that captures emotions of the other and generates an emotion distribution with multi decoders. Each decoder is optimized to deal with certain emotions and generate an empathetic response through softly combining the output emotion distribution.

- **MIME** (Majumder et al., 2020): Another Transformer-based model with the notion of mimicing the emotion of the other to a varying degree by group emotions into two clusters. It also introduces stochasticity to yield emotionally more varied empathetic responses.

- **BlenderBot-Joint** (Liu et al., 2021): A strong baseline model on the ESConv dataset, which prepends the special strategy token at the beginning of responses and conditions the generation process on it.

- **GLHG** (Peng et al., 2022): A hierarchical graph neural network to model the relationships between the global user's emotion causes and the local intentions for emotional support dialogue generation.

- **MISC** (Tu et al., 2022): An encoder-decoder model that leverages external commonsense knowledge to infer the seeker's fine-grained emotional status and respond skillfully using a mixture of strategy.