Knowledge-enhanced Memory Model for Emotional Support Conversation

Mengzhao Jia¹, Qianglong Chen², Liqiang Jing³, Dawei Fu⁴ and Renyu Li⁴

¹Shandong University, China
²Zhejiang University, China
³University of Texas at Dallas, TX, USA
⁴Alibaba Group, China

Abstract

The prevalence of mental disorders has become a significant issue, leading to the increased focus on Emotional Support Conversation as an effective supplement for mental health support. Existing methods have achieved compelling results, however, they still face three challenges: 1) variability of emotions, 2) practicality of the response, and 3) intricate strategy modeling. To address these challenges, we propose a novel knowledge-enhanced Memory mODEl for emotional suppoRt coNversation (MODERN). Specifically, we first devise a knowledge-enriched dialogue context encoding to perceive the dynamic emotion change of different periods of the conversation for coherent user state modeling and select context-related concepts from ConceptNet for practical response generation. Thereafter, we implement a novel memory-enhanced strategy modeling module to model the semantic patterns behind the strategy categories. Extensive experiments on a widely used large-scale dataset verify the superiority of our model over cutting-edge baselines.

Keywords

Mental disorders, Emotional Support Conversation, Mental health support, ConceptNet

1. Introduction

Mental disorders are known for their high burden, with more than 50% of adults experiencing a mental illness or disorder at some point in their lives; yet despite its high prevalence, only one in five patients receive professional treatment¹. Recent studies have shown that an effective mental health intervention method is the provision of emotional support conversations [1, 2]. As such, Emotional Support Conversations (ESConv), as defined by Liu et al. [3], has garnered substantial attention in recent years. They have emerged as a promising alternative strategy for mental health intervention, paving the way for the development of neural dialogue systems designed to provide support for those in need.

The ESConv takes place between a help-seeker (user) and a supporter (dialogue model) in a multi-turn manner. It requires the dialogue model to employ a range of supportive strategies effectively, easing the emotional distress of the users and helping them overcome the challenges they face. Prior research primarily concentrated on two aspects. The first aimed to enhance the model’s comprehension of the contextual semantics in the conversations, such as the user’s situation, emotions, and intentions. An example of these efforts is the work of Peng et al. [4] who designed a hierarchical graph network to capture the overall emotional problem cause and specific user intentions. The second aspect focused on predicting the dialogue strategy accurately and responding based on the predicted strategy category. For example, Cheng et al. [5] employed a lookahead heuristics for dialogue strategy planning and selection.

Despite the success of existing studies, this task is non-trivial due to the following three challenges.

- **Variability of emotions.** As the conversation progresses, the user’s emotional state evolves subtly and constantly. Accurately recognizing the emotional change is indispensable to understanding the user’s real-time state and thus responding empathically [7, 8, 9]. How to model the dynamic emotional change during the dialogue process is the first challenge.

- **Practicality of the response.** In the absence of explicit cues, neural dialogue systems are inclined to make generic responses [10, 11]. As shown in Figure 1, the generic responses are deficient to provide personalized and suitable suggestions for the user’s specific concerns. To resolve this issue, introducing context-related concepts (doctor and recover) can promote generating more meaningful and actionable suggestions for specific situations. As a result, the integration of appropriate concepts poses a non-trivial challenge in generating practical responses.
• **Intricate strategy modeling.** Dialogue strategy, as a kind of linguistic pattern, has been reported as a highly complex concept encompassing various intricate linguistic features [12, 13]. Previous work resorted to a single vector (a category indicator) for strategy representation, which is insufficient to fully represent the complex strategy pattern information. Therefore, how to model strategy information sufficiently is the third challenge.

To overcome these challenges, as shown in Figure 2, we introduce a novel knowledge-enhanced Memory mODEl for emotional suppoRt coNversation, dubbed MODERN. In particular, MODERN adapts the BART [14] as its backbone and consists of the knowledge-enriched dialogue context encoding module, the memory-enhanced strategy modeling module, and the response decoding module. To capture the emotional change as the conversation progresses, the first module detects the emotions for all utterances and explicitly injects them into the dialogue context as a kind of emotional knowledge, thus understanding the user’s status coherently. In addition, this module also introduces the concepts reasoning and selection component to acquire valid context-related concepts from a knowledge graph called ConceptNet [6] and incorporate them into the dialogue context to fulfill meaningful and practical suggestion generation. Moreover, in contrast to existing studies that depend on simplistic indicators to represent strategy categories, the memory-enhanced strategy modeling module learns strategy patterns by a strategy-specific memory bank. In this way, it can detect and mimic the intricate patterns in human emotional support strategies. Finally, the third module aims to generate the target response with the BART decoder.

Our contributions can be summarized as follows: 1) We first analyze the current challenges of the ESConv task, and according to that propose a novel knowledge-enhanced Memory mODEl for emotional suppoRt coNversation, named MODERN, which can model complex supportive strategy as well as utilize emotional knowledge and context-related concepts to perceive the variability of emotions and provide practical support advice. 2) We propose a memory-enhanced strategy modeling module, where a unique memory bank is designed to model intricate strategy patterns, and an auxiliary strategy classification task is introduced to distill the strategy pattern information. 3) We present a thorough validation and evaluation of our model, providing an in-depth analysis of the results and a comparison with other models. The extensive experiments on the ESConv dataset [3] demonstrate that MODERN achieves state-of-the-art performance under both automatic and human evaluations. The code is available at https://projs2release.wixsite.com/modern.

### 2. Related Work

#### 2.1. Emotional and Empathetic Dialogue Systems

With the popularity and growing success of dialogue systems, many research interests have recently endeavored to empower the system to reply with a specific and proper emotion, therefore forming a more human-like conversation. Particularly, two research directions arise researchers’ interest, namely the emotional [15] and empathetic [16] response generation. The former direction expects the dialogue agent to respond with a given emotion [17, 18, 19, 20]. While the latter requires the dialogue system to actively detect and understand the users’ emotions and then respond with an appropriate emotion [21]. For example, Lin et al. utilized multiple decoders as different listeners to react to different emotions and then softly combine the output states of the decoders appropriately based on the recognized user’s emotion. Nevertheless, unlike above directions, the ESConv task concentrates on alleviating users’ negative emotion intensity and providing supportive instructions to help them overcome struggling situations.

#### 2.2. Emotional Support Conversation

As an emerging research task, emotional support conversation has gradually attracted intense attention in recent years. Existing works mostly focus on two aspects. The first is to understand the complicated user emotions and
intentions in the dialogue context. Specifically, they explored context semantic relationships [23, 4], commonsense knowledge [24, 4], or emotion causes [5] to better capture and understand the emotions and intentions of users. The other trend in addressing the task is to predict the strategy category accurately so as to respond in accordance with it [25, 23]. For instance, Tu et al. firstly proposed to predict a strategy probability distribution and generate the response guided by a mixture of multiple strategies. Despite their remarkable achievements, existing work still face three challenges: intricate strategy modeling, variability of emotions, and practicality of the responses.

3. Task Formulation

For concise mathematical expression, we first declare some notations in this paper. We use bold capital letters \( \mathbf{X} \) and bold lowercase letters \( \mathbf{x} \) to represent matrices and vectors, respectively. We adopt non-bold letters \( (N) \) to denote scalars. Greek letters \( (\lambda) \) refer to hyperparameters. All the vectors, if not clarified, are in column forms.

In the setting of emotional support conversation, the dialogue is participated by a help-seeker and a supporter, and the latter tries to comfort and support the help-seeker to lower he/she’s emotional intensity level. The target is to generate a response based on the dialogue history. Besides, the supporter is required to select one of \( G \) strategies and respond accordingly. Suppose we have a training dataset \( \mathcal{P} = \{ p_1, p_2, \ldots, p_V \} \) composed of \( V \) samples. Each sample \( p_i = \{ t_i, \mathcal{D}_i, g_i, R_i \} \) includes a seeker’s situation \( t_i \), a dialogue context \( \mathcal{D}_i \), a support strategy \( g_i \), and a target response \( R_i \). Therein, \( \mathcal{D}_i \) contains a sequence of \( N_i \) history utterances between the user and the supporter, denoted as \( \mathcal{D}_i = (u_1, u_2, \ldots, u_{N_i}) \). \( R_i = (r_1, r_2, \ldots, r_{N_i}) \) is the supportive response with \( N_i \) tokens. The goal of the ESConv task is to learn a model \( \mathcal{F} \) that can generate a supportive response \( \hat{R}_i \) referring to the input context \( \mathcal{D}_i \) and situation \( t_i \) as follows,

\[
\hat{R}_i = \mathcal{F}(\mathcal{D}_i, t_i|\Theta),
\]

where \( \Theta \) refers to the set of to-be-learned parameters of the model \( \mathcal{F} \). Notably, the ground truth \( g_i \) can be utilized in the model training stage but is not available and need to be predicted in the inference phrase. For brevity, we temporarily omit the superscript \( i \) of the \( i \)-th sample in the rest of this paper.

4. Method

In this section, we detail the proposed model MODERN, which comprises three main components: knowledge-enriched dialogue context encoding, memory-enhanced strategy modeling, and response decoding, demonstrated in Figure 2.

4.1. Knowledge-enriched Dialogue Context Encoding

In this section, we first utilize an emotion detector to recognize fine-grained emotions of utterances as emotional knowledge for capturing emotional change. In addition, we select related concepts from the ConceptNet knowledge graph for meaningful and practical suggestion generation.
4.1.1. Change-aware Emotion Detection

Psychiatric and mental health studies have proved that empathy is essential to emotionally helping relationships [26, 27, 28]. And fine-grained emotional information is one of the key factors to enhance the empathetic ability [29]. Apart from the static emotional signals, dynamic emotional changes during the conversation progress are also beneficial. Concretely, change-aware emotion information enriches the model to understand the user’s status coherently. Inspired by this, we devise to identify the user’s fine-grained emotions and perceive the dynamic changes of emotions in the dialogue context.

Specifically, we start by obtaining the fine-grained emotion via an off-the-shelf pretrained emotion detector⁵, which can recognize up to 28 different emotional categories. We apply the detector to every utterance in the dialogue context for emotion recognition as follows,

\[ e_j = \text{Emo}(u_j), \quad 0 \leq j \leq N_u \]  

where \( e_j \) is the predicted emotional category word representing the detected emotion in \( u_j \). Thereafter, we directly inject the natural language form of emotion category words into the dialogue context as additional emotion knowledge. This practice aligns closely with the input format of the pretrained BART model. Moreover, it also avoids introducing unnecessary parameters that would interfere with the generative model learning. Empirically, we concatenate the emotional category into the sequence of dialogue context tokens, denoted as \( I = [u_1; e_1; \text{SEP}; ...; u_J; e_J; \text{SEP}; ...; u_N; e_N] \). In this way, the dynamic emotional changes corresponding to the dialogue progress can be coherently exploit by the emotional support model.

4.1.2. Context-related Concepts Reasoning and Selection

Considering ConceptNet involve abundant general human knowledge, which plays an important role in understanding human situations and associating them with practical suggestions, we select potentially useful context-related concepts to enrich the model to generate responses with high informativeness. For instance, the commonsense knowledge database can easily relate failing exam to academic stress. In terms of daily human activities, this knowledge serves as a guide for dealing with daily affairs and problems. Such knowledge is useful for providing advice and guidance in the emotional support system. Therefore, we mine and associate commonsense knowledge of the data to provide potentially useful information for generating instructive responses. Concretely, the ConceptNet knowledge graph involves 3.1 million concepts and 38 million relations, which can be used to mine the underlying concept information in the dialogue context. We first identify all the concepts in ConceptNet that are mentioned in the dialogue context and remove the top-K frequent concepts in the training set because these words are usually too general to provide valid suggestions for a specific situation, such as help, thing, and feeling. In this way, we derive \( N_c \) concepts, denoted as \( \{c_1, ..., c_{N_c}\} \).

Thereafter, we leverage the \( N_c \) concepts as the anchors to reason the related concepts. Specifically, for each anchor concept \( c \), we retrieve all its one-hop neighboring concept-relation pairs from the ConceptNet. Mathematically, let \( \mathcal{A}(c) \) be the set of neighboring concept-relation pairs of the anchor concept \( c \). Then the context-related concept-relation pairs can be represented as \( \{\mathcal{A}(c_1), \mathcal{A}(c_2), ..., \mathcal{A}(c_{N_c})\} \), where \( \mathcal{A}(c_i) = \{(a_{i1}, h_{i1}), (a_{i2}, h_{i2}), ..., (a_{iN_i}, h_{iN_i})\} \) is a set of \( N_i \) retrieved \(<\text{concept, relation}>\) pairs. Following Liu et al. [30], to avoid introducing extraneous concepts, we filter out concepts with excluded relations, Antonym, ExternalURL, and NotCapableOf. Finally, we concatenate the concepts in the filtered pairs and deem them as context-related concepts \( C \).

4.1.3. Knowledge-enriched Dialogue Context Embedding

To acquire the representation of the dialogue context and corresponding knowledge, we first encode the dialogue context sequence (user situation \( t \), emotional-aware dialogue context \( I \), and concepts \( C \)) with a Transformer-based encoder as follows,

\[ H = \text{Enc}_\text{d}([f; I; C]), \]  

where \( H \in \mathbb{R}^{L \times d} \) denotes the hidden contextual representation. \( L \) and \( d \) refer to the number of tokens in \([f; I; C]\) and the representation hidden dimension, respectively.

4.2. Memory-enhanced Strategy Modeling

During the conversation, the supporter adopts different strategies for different purposes, ultimately achieving the goal of reducing the intensity of the user’s negative emotions. For example, using the Question strategy helps the supporter to explore the concrete situation faced by the user, while the use of the Reflection of Feelings strategy conveys the supporter’s understanding of the user’s current emotions. Existing work constrains the model to responding under a strategy category by simply providing a single vector indicating the strategy’s name or description. However, the semantic patterns of strategies are highly complex, the name or description is not able to contain the specific linguistic patterns (expression manner, wording, and phrasing) of the strategy. Therefore,
inspired by [31], we propose to disentangle the strategy patterns from the same-strategy responses to provide more specific guidance for the strategy-constrained generation.

4.2.1. Strategy Pattern Modeling

We first acquire strategy pattern representations of each response in the training set via a strategy pattern extractor $Enc$, as follows,

$$r = \text{MaxPooling}(Enc(R)),$$

where $r \in \mathbb{R}^d$ denotes the strategy pattern representation. Meanwhile, in order to accurately capture the strategy pattern information and avoid irrelevant disturbance, we design an auxiliary strategy classification task to guide the extractor to map more strategy-related information into the pattern representation. The auxiliary objective $\mathcal{L}_r$ is derived by the following loss function,

$$\mathcal{L}_r = -\log p(g|r).$$

4.2.2. Strategy-specific Memory Bank

To utilize detail and ample strategy pattern information, instead of using a single representation vector, we devise a memory bank mechanism to store multiple pattern representations according to their strategy categories. In particular, we denote the memory bank as $\mathcal{M} = \{M^1, ..., M^G\}$, in which $M^g \in \mathbb{R}^{N^g \times d}$ is a matrix of $N^g$ pattern representations corresponding to $g$-th strategy category, and $G$ is the total number of strategy categories. $N^g_0$ is 0 at the initial training step. As the training progresses, $N^g$ continues to increase until the maximum threshold $N_m$ is reached. In particular, we store pattern representation of $g$-th strategy category into the corresponding $M^g$ by concatenation as follows,

$$M^g \leftarrow [M^g; r_{g}],$$

where $r_g$ denotes a representation belongs to the $g$-th strategy category and $\left[ : \right]$ refers to the concatenation operation. As the representations are optimized along with the classification training process, we dynamically update $M^g$ in a first-in-first-out manner as follows,

$$M^g = \begin{cases} M^g_{\left[ \left[N^g_0 - N_m, N^g_0 \right] \right]} & \text{if } N^g > N_m \\ M^g, & \text{otherwise} \end{cases}$$

where $N_m$ and $N^g_0$ denote the maximum storage limit and the current storage volume of each memory matrix, respectively. The algorithm of the memory storing and updating operation is summarized in the appendix.

4.2.3. Strategy Prediction

In order to use the strategy pattern information in the memory bank, the model requires selecting a proper strategy category based on the dialogue context. To achieve this, we leverage a strategy predictor, which aims to capture information relevant to strategy decisions in the context.

The strategy predictor is composed of a Transformer-based encoder and a classification module. The encoder first encodes the dialogue context into a strategy predict representation. It is worth noting that we adopt independent representations for strategy prediction and response generation tasks considering the fact that jointly optimal solutions may not always exist for different tasks. Subsequently, the classification module maps the vector as a $G$ dimension vector, which is regarded as the probability distribution of the $G$ strategy types. Formally, the strategy prediction can be written as follows,

$$\{s = \text{MaxPooling}(Enc(I)), \hat{g} = \text{argmax}(\sigma(\text{MLP}(s)))\},$$

where $Enc$ is a Transformer-based encoder. The strategy prediction representation is denoted as $s \in \mathbb{R}^d$. MLP and $\sigma()$ are a multi-layer perceptron and the Sigmoid function, respectively. The argmax operation is used to obtain the predicted strategy category $\hat{g}$. We use the following objective to optimize the strategy prediction task,

$$\mathcal{L}_s = -\log p(g|s).$$

4.2.4. Memory-enhanced Encoding

After predicting the strategy category $\hat{g}$, instead of directly using $\hat{g}$ as an indicator to constrain the response generation, we integrate the aforementioned corresponding memory bank matrix $M^\hat{g}$ and the context representation, so as to fully exploit the abundant pattern information of the particular strategy.

Empirically, we fuse the matrix and the context representation with a multi-head cross-attention module [32] as follows,

$$m = \text{MaxPooling}(\text{CrossAtt}(H, M^\hat{g})).$$

where $H$ and $M^\hat{g}$ act as the query and the key-value pair in the cross-attention, respectively, $m \in \mathbb{R}^d$ denotes the memory-enhanced strategy modeling feature.

4.3. Response Decoding

In order to generate the emotional supportive response, we input the encoded features, the memory-enhanced strategy modeling feature $m$ and the knowledge-enriched dialogue context embedding $H$, into the Transformer
We conducted experiments on the ESConv dataset [3].

\[ \mathcal{L} = \mathcal{L}_g + \lambda_1 \mathcal{L}_i + \lambda_2 \mathcal{L}_n, \]

where \( \mathcal{L} = \mathcal{L}_g \) refers to the previous generated \( l - 1 \) tokens of the target response. \( \mathcal{L}_g \) denotes the decoder module. Notably, to avoid the accumulated error, we replace \( \mathcal{L}_g \) by \( r_{cd} \) in the training phase. For optimization, we introduce the standard cross-entropy loss function for response generation as follows,

\[ \mathcal{L}_g = - \frac{1}{N_m} \sum_{i=1}^{N_m} \log p(r_i | r_{cd}, E), \]

where \( N_m \) denotes the length of the target response.

5. Experiments

5.1. Dataset

We conducted experiments on the ESCConv dataset [3]. Each sample in the dataset is a dialogue between a help-seeker and a supporter. In addition to the context, it also contains rich information, the situation that the help-seeker’s is facing. It also provides the annotation of the strategy category used in every supporter’s response. The dataset contains 1,300 long conversations and overall 38,350 utterances, with an average of 29.5 utterances in each dialogue. For the data split, we followed the same setting of previous work [3, 5].

5.2. Implementation Details

Following the setting of the previous study [5], we also utilized encoder and decoder of the pretrained BART\(^3\) provided by HuggingFace [33] to initialize the parameters of the context encoder, the strategy predictor, and the decoder, respectively. The number of layers in the encoder and decoder are both 6. The dimension of hidden feature \( d \) equals to 512. To form a mini-batch, the input sequence length \( L \) is unified to 512. The hidden dimension \( d \) is 768. The category number \( G \) equals 8. The maximum memory storage \( N_m \) is set as 64. \( K \) equals 20. \( \lambda_1 \) and \( \lambda_2 \) are 0.3 and 0.1, respectively. The batch size is 16. We use the PPL metric on the validation set to monitor the training progress.

5.3. Evaluation Metrics

For the comprehensive evaluation, we conducted both automatic and human evaluations.

**Automatic Evaluation** For the automatic evaluation, we adopt several mainstream metrics commonly used in dialogue response tasks, PPL (Perplexity), BLEU-\{1,2,3,4\} (B-{1,2,3,4}) [36], ROUGE-L (R-L) [37], METEOR (MIT) [38], and CIDEr [39].

**Human Evaluation** Apart from the automatic evaluation, we also consider the human evaluation, as it has been reported [40] that the automatic evaluation is sometimes unreliable in generation tasks. We first randomly select 70 testing samples for evaluation. Then, we employ 2 volunteers to manually choose which response outperforms the other one. Every sample is annotated twice. For each case, the volunteers need to compare the generated texts of different models according to the following four dimensions: 1) Fluency: which response

---

3https://huggingface.co/facebook/bart-base.
Table 1
Performance comparison under automatic evaluations. The best results are highlighted in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL</th>
<th>B-1</th>
<th>B-2</th>
<th>B-3</th>
<th>B-4</th>
<th>R-L</th>
<th>MT</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoEL</td>
<td>264.11</td>
<td>19.04</td>
<td>6.47</td>
<td>2.91</td>
<td>1.51</td>
<td>15.95</td>
<td>7.96</td>
<td>10.95</td>
</tr>
<tr>
<td>MIME</td>
<td>69.28</td>
<td>15.24</td>
<td>5.56</td>
<td>2.64</td>
<td>1.50</td>
<td>16.12</td>
<td>6.43</td>
<td>10.66</td>
</tr>
<tr>
<td>DialoGPT-Joint</td>
<td>15.71</td>
<td>17.39</td>
<td>5.59</td>
<td>2.03</td>
<td>1.18</td>
<td>17.91</td>
<td>7.55</td>
<td>11.86</td>
</tr>
<tr>
<td>BlenderBot-Joint</td>
<td>16.79</td>
<td>17.62</td>
<td>6.91</td>
<td>2.81</td>
<td>1.66</td>
<td>17.94</td>
<td>7.54</td>
<td>18.04</td>
</tr>
<tr>
<td>MISC</td>
<td>16.16</td>
<td>-</td>
<td>7.31</td>
<td>-</td>
<td>2.20</td>
<td>17.91</td>
<td>11.05</td>
<td>-</td>
</tr>
<tr>
<td>GLHG</td>
<td>15.67</td>
<td>19.66</td>
<td>7.57</td>
<td>3.74</td>
<td>2.13</td>
<td>16.37</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FADO</td>
<td>15.72</td>
<td>-</td>
<td>8.00</td>
<td>4.00</td>
<td>2.32</td>
<td>17.53</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PoKE</td>
<td>15.84</td>
<td>18.41</td>
<td>6.79</td>
<td>3.24</td>
<td>1.78</td>
<td>15.84</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MultiESC</td>
<td>13.41</td>
<td>21.65</td>
<td>9.18</td>
<td>4.99</td>
<td>3.09</td>
<td>20.41</td>
<td>8.84</td>
<td>30.08</td>
</tr>
</tbody>
</table>

Dialogue Context
Supporter: Good Afternoon. How are you doing today?
User: I am Ok thanks but have an unusual issue. I think my girlfriend may be cheating on me but I'm too scared to do anything about it as I don't want her to leave me. I feel ashamed of myself but I can't help it.

Ground Truth: (Self-disclosure) I've dealt with the same problem with my partner more than once. I love him very much and I found him to be a 9 to my 5 as well.

w/o-Mem: (Reflection of Feelings) I am sorry to hear that. I am sorry you are feeling that way. I know how you feel.

MODERN: (Self-disclosure) I can understand how that would be difficult. I had a similar situation with my ex-boyfriend. I know how hard it can be to let go of someone.

(a)

Dialogue Context
Supporter: Hello, what can I help you with this evening?
User: I am feeling very depressed lately. Like a constant pressure that I don't know exactly what it is.

Ground Truth: (Self-disclosure) I believe there are special lamps that you could get to help. I personally have been trying to get up a little earlier in the morning and enjoy my coffee with the sunrise...

w/o-KG: (Providing Suggestions) I've been in that situation myself, and I've found that it's very easy to get depressed.

MODERN: (Providing Suggestions) I've found that taking a walk or sitting down to write out a list of things that you'd like to do helps to clear your mind.

(b)

Table 2
The human evaluation results in four dimensions.

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>Aspects</th>
<th>Win</th>
<th>Tie</th>
<th>Lose</th>
</tr>
</thead>
<tbody>
<tr>
<td>vs. FADO</td>
<td>Fluency</td>
<td>37.0</td>
<td>24.0</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>Relevance</td>
<td>35.5</td>
<td>20.5</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Empathy</td>
<td>34.0</td>
<td>16.5</td>
<td>19.5</td>
</tr>
<tr>
<td></td>
<td>Information</td>
<td>37.5</td>
<td>17.0</td>
<td>15.5</td>
</tr>
<tr>
<td>vs. MultiESC</td>
<td>Fluency</td>
<td>30.5</td>
<td>24.5</td>
<td>15.0</td>
</tr>
<tr>
<td></td>
<td>Relevance</td>
<td>32.5</td>
<td>20.5</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td>Empathy</td>
<td>35.0</td>
<td>19.5</td>
<td>15.5</td>
</tr>
<tr>
<td></td>
<td>Information</td>
<td>37.5</td>
<td>21.0</td>
<td>11.5</td>
</tr>
</tbody>
</table>

is more fluent, correct, and coherent in grammar and syntax; 2) Relevance: which response talks more relevantly regarding current dialogue context; 3) Empathy: which response is better to react with appropriate emotion according to the user’s emotional state; 4) Information: which response provides more suggestive information to help solve the problem. To further control the quality of the evaluation, we also invite an inspector to randomly sample and double-check 10% rating results.

5.4. Model Comparison & Analysis
To validate the effectiveness of our model, we compare it with several state-of-the-art baselines:

- **MoEL** [22]. This method adopts multiple decoders as different listeners for different emotions. The outputs of decoders are softly combined to generate the response.
- **MIME** [21]. This model shares the same architecture as MoEL and extends it to mimic the speaker’s emotion.
- **DialoGPT-Joint** [3]. This model is built on a pre-trained dialog agent DialoGPT [41]. It first predicts a strategy and prepends a special token before the response sentence to control the generation under that strategy.
- **BlenderBot-Joint** [3]. This model adopts the same strategy prediction and generation scheme as DialoGPT. Differ from the former one, it is built on a pre-trained conversational response generation model named BlenderBot [42].
- **MISC** [24]. This model also adopts BlenderBot as the backbone. It leverages common sense
knowledge to enhance the understanding of the speaker’s emotional state, and the response is generated conditioned on a mixture of strategy distribution.

- **GLHG** [4]. This model has a global-to-local hierarchical graph structure. It models the global cause and the local intention of the speaker to provide more supportive responses.
- **PoKE** [23]. This work utilized Conditional Variational Autoencoder [43] to model the mixed strategy.
- **FADO** [25]. This work devises a dual-level feedback strategy selector to encourage or penalize strategies during the strategy selection process.
- **MultiESC** [5]. This work proposes lookahead heuristics to estimate the future strategy and capture users’ subtle emotional expressions with the NRC VAD lexicon [44] for user state modeling.

### Automatic Evaluation

We compared our model with the above baselines using automatic metrics and results are reported in Table 1. As we can see, 1) MODERN outperforms the baselines in most metrics, which is a powerful proof of the effectiveness of the proposed method. 2) The models with BART backbone (MultiESC and MODERN) surpass those baselines with BlenderBot [42] backbone (BlenderBot-Joint, MISC, GLHG, FADO, and PoKE) across most of the metrics despite the latter being pretrained on empathetic-related data. One possible explanation is that the ESConv task requires sophisticated linguistic knowledge (e.g. correct grammar and wording appropriate to the current situation) in addition to empathetic ability. 3) Our MODERN consistently exceeds MultiESC with the same BART backbone. This suggests that BART model with large-scale pretraining still requires strategy pattern information and knowledge (emotion and concepts) to further facilitate supportive response generation.

### Human Evaluation

For human evaluation, we report the comparison results between our model and the two best baselines (FADO and MultiESC) in Table 2. In particular, for each pair of model comparisons and each metric, we show the number of samples where our model achieves better (denoted as "Win"), equal (denoted as "Tie"), and worse performance (denoted as "Lose") compared with the baselines. As seen, MODERN outperforms all baselines across different evaluation metrics, as the number of "Win" cases is always significantly larger than that of "Lose" cases in each pair of model comparisons, which is consistent with the results in Table 1. In addition, the number of "Win" cases is the largest for the Information metric compared with other metrics, which demonstrates that integrating context-related concepts can supply meaningful information for emotional support.

### 5.5. Ablation Study

We compare the original MODERN model with the following derivatives to demonstrate that all the designed modules are essential for the ESConv task. 1) **w/o-ℒ₁**. To show the benefit of constraint for strategy prediction, we removed the corresponding loss function by setting $\lambda_2 = 0$ in Eqn.(13). 2) **w/o-ℒ₂**. To show the effect of the auxiliary strategy classification task, we removed the corresponding loss function by setting $\lambda_2 = 0$ in Eqn.(13). 3) **w/o-Mem**. In this derivative, we disabled the memory module, which stores pattern representations of different strategies. 4) **w/o-Emo**. In this derivative, we removed the change-aware emotion detection module. And 5) **w/o-KG**. We discarded the context-relate concepts reasoning and selection component.

We provided the ablation study results on the ESConv dataset in Table 3 in terms of all metrics. From this table, we have the following observations: 1) Our MODERN consistently outperforms w/o-Mem, especially on the BLUE metrics (B-{1,2,3,4}), which suggests that the memory-enhanced strategy modeling module can provide sufficient linguistic pattern references and hence boost the performance of generating responses in accordance with specific strategy categories. 2) MODERN exceeds w/o-ℒ₂ across all metrics. This verifies it is indispensable to constrain the model to predict and respond under proper strategies. 3) w/o-ℒ₁ obtains a slightly worse result than MODERN, which demonstrates the strategy classification auxiliary task indeed helps with guiding the pattern representation learning. And 4) w/o-Emo and w/o-KG both perform worse than MODERN, which demonstrates the importance of change-aware emotion and context-related concepts. Notably, w/o-KG surpasses w/o-Emo. One possible explanation is that being aware of the dynamic emotional changes during the conversation facilitates the model to provide empathy and emotional support accordingly.
5.6. Case Study
We illustrate several conversations in the test set to get an intuitive understanding of our model in Figure 3. We showed two samples and compare the responses generated by MODERN and two derivatives, w/o-Mem and w/o-KG. As can be seen in case (a), MODERN fulfills to respond with the strategy of Self-disclosure and generates a contextually appropriate response. Being equipped with a memory-enhanced strategy modeling module, MODERN shares a similar experience closely related to the seeker’s problem relationship issue. Whereas the w/o-Mem model generates a plain and monotonous response, which is not very relevant to the user’s current issue. The other case (b) demonstrates how MODERN reaps benefits from external knowledge. Based on the situation that the seeker mentions feeling depressed, MODERN leverages the context-related concepts and associates this emotion status with practical suggestions taking a walk or sitting down to write... effectively. While without relevant knowledge, the response generated by the w/o-KG derivative is relatively vague and less specific, which is deficient to benefit the user’s situation.

6. Conclusions
In this paper, we propose a novel knowledge-enhanced Memory mODEl for emotional support conversation, dubbed MODERN, which can perceive fine-grained emotional changes in the conversation, utilizes the concepts from knowledge graph to facilitate generating responses with practical suggestions, and model concrete strategy semantic patterns with memory bank mechanism. Both automatic and human evaluation results show that our model surpasses the state-of-the-art methods in emotional support conversation. In addition, the ablation study demonstrates the effectiveness of each component of our model.

Limitations
The ESConv task requires the dialogue agent to reveal some information about itself. For example, one of the strategies called Self-disclosure, expects the agent to cite their own experience. However, in our experiments, we observed that the current model often struggles to maintain a consistent personality. We speculate that this may be due to the fact that the supporter role in the full training sample is provided by multiple individuals, and thus there is no uniform character experience and story, which leads to the problem of inconsistent personal experiences during the conversation. We believe that how to make the dialogue agent show coherent and unified personal information and experiences in the ESConv task deserves the attention of future work.

Ethical Considerations
The dataset used in our work is a publicly available dataset that has been widely used in the field of emotional support conversation. Sensitive and personally identifiable information was filtered during the construction of the dataset. In the work of this paper, our model focuses on informal, emotional support provided between friends’ daily chats and does not provide professional mental health diagnosis and counseling services. The use of this model should be avoided for patients with serious mental illnesses, such as self-harm-related conversations, in order to prevent triggering serious consequences.

References
[27] B. Liu, S. S. Sundar, Should machines express sympathy and empathy? experiments with a health advice chatbot, Cyberpsychology, Behavior, and