Knowledge-enhanced Memory Model for Emotional Support Conversation*

Mengzhao Jia[1], Qianglong Chen[2], Liqiang Jing[3], Dawei Fu[4] and Renyu Li[4,*]

[1]*Shandong University, China*
[2]*Zhejiang University, China*
[3]*University of Texas at Dallas, TX, USA*
[4]*Alibaba 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 treatment[1]. 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.

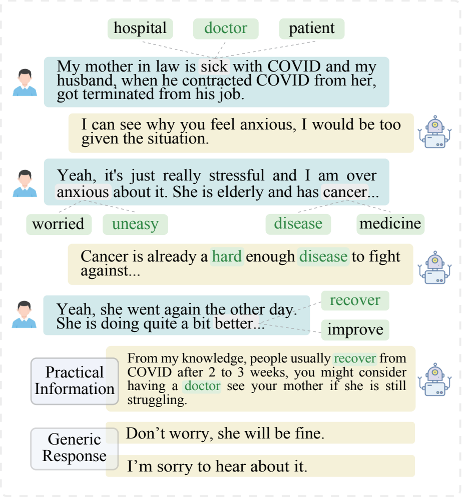
- **Variability of emotions.** As the conversation progresses, the user's emotional state evolves subtly and constantly. Accurately recognizing the emotional change is indispensable to understanding the user's real-time state and thus responding empathically [7, 8, 9]. How to model the dynamic emotional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

Machine Learning for Cognitive and Mental Health Workshop (ML4CMH), AAAI 2024, Vancouver, BC, Canada

*Corresponding author.

jiamengzhao98@gmail.com (M. Jia); chenqianglong@zju.edu.cn (Q. Chen); jingliqiang6@gmail.com (L. Jing); dawei.fdw@alibaba-inc.com (D. Fu); renyu.rl@alibaba-inc.com (R. Li)

[1] https://tinyurl.com/4r8svsst.



My mother in law is sick with COVID and my husband, when he contracted COVID from her, got terminated from his job.

I can see why you feel anxious, I would be too given the situation.

Yeah, it's just really stressful and I am over anxious about it. She is elderly and has cancer...

Cancer is already a hard enough disease to fight against...

Yeah, she went again the other day. She is doing quite a bit better...

From my knowledge, people usually recover from COVID after 2 to 3 weeks, you might consider having a doctor see your mother if she is still struggling.

Don't worry, she will be fine.

I'm sorry to hear about it.

Figure 1: Illustration of an emotional support conversation example. The words with a green background are the retrieved concepts from ConceptNet [6].

- **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 avaliable 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

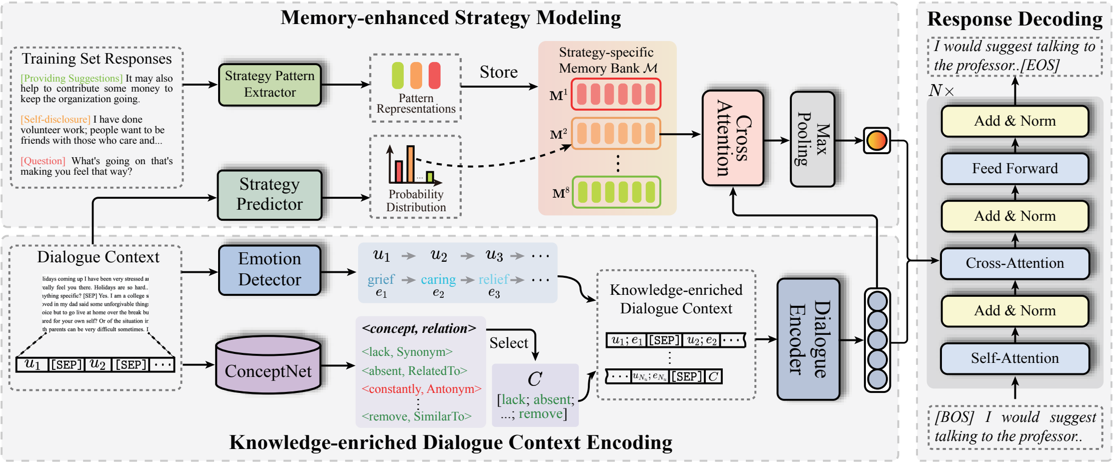

Memory-enhanced Strategy Modeling
Training Set Responses
[Providing Suggestions] It may also help to contribute some money to keep the organization going.
[Self-disclosure] I have done volunteer work; people want to be friends with those who care and...
[Question] What's going on that's making you feel that way?
Strategy Pattern Extractor
Pattern Representations
Store
Strategy-specific Memory Bank $\mathcal{M}$
$\mathbf{M}^1$
$\mathbf{M}^2$
$\mathbf{M}^8$
Strategy Predictor
Probability Distribution
Cross Attention
Max Pooling
Dialogue Context
u_1 [SEP] u_2 [SEP] $\cdots$
Emotion Detector
u_1 → u_2 → u_3 → $\cdots$
grief → caring → relief → $\cdots$
e_1 e_2 e_3
ConceptNet
<concept, relation>
<lack, Synonym>
<absent, RelatedTo>
<constantly, Antonym>
<remove, SimilarTo>
Select
C
[lack; absent; ...; remove]
Knowledge-enriched Dialogue Context
$u_1; e_1$ [SEP] $u_2; e_2$ $\cdots$
$\cdots$ $u_{N_u}; e_{N_u}$ [SEP] C
Dialogue Encoder
Knowledge-enriched Dialogue Context Encoding
Response Decoding
I would suggest talking to the professor..[EOS]
$N\times$
Add & Norm
Feed Forward
Add & Norm
Cross-Attention
Add & Norm
Self-Attention
[BOS] I would suggest talking to the professor..

Figure 2: Illustration of the proposed MODERN framework, which consists of three key components: Strategy Memory-enhanced Dialogue Context Encoding, Multi-source Knowledge Injection, and Response Decoding.

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 (λ) 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 $\mathscr{P} = \{p_1, p_2, \dots, p_V\}$ composed of V samples. Each sample $p_i = \{t_i, \mathscr{D}_i, g_i, R_i\}$ includes a seeker's situation t_i, a dialogue context $\mathscr{D}_i$, a support strategy g_i, and a target response R_i. Therein, $\mathscr{D}_i$ contains a sequence of N_u^i history utterances between the user and the supporter, denoted as $\mathscr{D}_i = (u_1^i, u_2^i, \dots, u_{N_u^i}^i)$. $R_i = (r_1^i, r_2^i, \dots, r_{N_r^i}^i)$ is the supportive response with N_r^i tokens. The goal of the ESConv task is to learn a model $\mathscr{F}$ that can generate a supportive response $\hat{R}_i$ referring to the input context $\mathscr{D}_i$ and situation t_i as follows,

$$\hat{R}_i = \mathscr{F}(\mathscr{D}_i, t_i|\Theta), \tag{1}$$

where Θ refers to the set of to-be-learned parameters of the model $\mathscr{F}$. 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 temporally 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 detector[2], 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,

$$e_j = \mathrm{Emo}(u_j), \quad 0 \leq j \leq N_u, \tag{2}$$

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 $I = [u_1; e_1; \mathrm{SEP}; \ldots u_j; e_j; \mathrm{SEP}; \ldots u_{N_u}; e_{N_u}]$. In this way, the dynamic emotional changes corresponding to the dialogue progress can be coherently exploit by the emotinal 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 $\{c_1, \cdots, c_{N_c}\}$.

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 $\mathcal{N}(c)$ be the set of neighboring concept-relation pairs of the anchor concept c. Then the context-related concept-relation pairs can be represented as $\{\mathcal{N}(c_1), \mathcal{N}(c_2), \cdots, \mathcal{N}(c_{N_c})\}$, where $\mathcal{N}(c_i) = \{(a_{c_1}^1, h_{c_1}^1), \cdots, (a_{c_1}^{N_{c_i}}, h_{c_1}^{N_{c_i}})\}$ is a set of N_{c_i} retrieved $< concept, relation >$ 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,

$$\mathbf{H} = \mathrm{Enc}_c([t; I; C]), \tag{3}$$

where $\mathbf{H} \in \mathbb{R}^{L \times d}$ denotes the hidden contextual representation. L and d refer to the number of tokens in $[t; I; C]$ 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,

[2] https://huggingface.co/arpanghoshal/EmoRoBERTa.

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 responses in the training set via a strategy pattern extractor Enc_r as follows,

$$\mathbf{r} = \text{MaxPooling}(\text{Enc}_r(R)), \tag{4}$$

where $\mathbf{r} \in \mathbb{R}^d$ 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,

$$\mathcal{L}_r = -\log p(g|\mathbf{r}). \tag{5}$$

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 $\mathcal{M} = \{\mathbf{M}^1, \ldots, \mathbf{M}^G\}$, in which $\mathbf{M}^g \in \mathbb{R}^{N_s^g \times d}$ is a matrix of N_s^g pattern representations corresponding to g-th strategy category, and G is the total number of strategy categories. N_s^g is 0 at the initial training step. As the training progresses, N_s^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 $\mathbf{M}^g$ by concatenation as follows,

$$\mathbf{M}^g \longleftarrow [\mathbf{M}^g; \mathbf{r}_g], \tag{6}$$

where $\mathbf{r}_g$ denotes a representation belongs to the g-th strategy category and $[\cdot;\cdot]$ refers to the concatenation operation. As the representations are optimized along with the classification training process, we dynamically update $\mathbf{M}^g$ in a *first-in-first-out* manner as follows,

$$\mathbf{M}^g = \begin{cases} \mathbf{M}^g_{[N_s^g - N_m : N_s^g]}, & \text{if } N_s^g > N_m \\ \mathbf{M}^g, & \text{otherwise} \end{cases} \tag{7}$$

where N_m and N_s^g 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 these, 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,

$$\begin{cases} \mathbf{s} = \text{MaxPooling}(\text{Enc}_s(I)), \\ \hat{g} = \text{argmax}(\sigma(\text{MLP}(\mathbf{s})), \end{cases} \tag{8}$$

where Enc_s is a Transformer-based encoder. The strategy prediction representation is denoted as $\mathbf{s} \in \mathbb{R}^d$. MLP and $\sigma(\cdot)$ 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,

$$\mathcal{L}_s = -\log p(g|\mathbf{s}). \tag{9}$$

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 $\mathbf{M}^g$ 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,

$$\mathbf{m} = \text{MaxPooling}(\text{CrossAtt}(\mathbf{H}, \mathbf{M}^g)), \tag{10}$$

where $\mathbf{H}$ and $\mathbf{M}^g$ act as the *query* and the *key-value* pair in the cross-attention, respectively, $\mathbf{m} \in \mathbb{R}^d$ 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 $\mathbf{m}$ and the knowledge-enriched dialogue context embedding $\mathbf{H}$, into the Transformer decoder. The generation process aims to predict the conditional probability distribution $p(\hat{r}_l|\hat{r}_{<l}, \mathbf{m}, \mathbf{H})$ in an auto-regressive manner, which means the decoder generates the l-th word conditioned on all previous generated words as well as the encoded representation. Formally, we deploy the decoding process, which predicts the conditional probability distribution over each token in the target response in an auto-regressive manner as follows,

$$p(\hat{r}_l \mid \hat{r}_{<l}, \mathbf{E}) = \mathrm{Dec}(\hat{r}_{<l}, \mathbf{E}), \tag{11}$$

where $\mathbf{E} = [\mathbf{m}; \mathbf{H}]$. $\hat{r}_{<l}$ refers to the previous generated $l-1$ tokens of the target response. $\mathrm{Dec}(\cdot)$ denotes the decoder module. Notably, to avoid the accumulated error, we replace $\hat{r}_{<l}$ by $r_{<l}$ in the training phase. For optimization, we introduce the standard cross-entropy loss function for response generation as follows,

$$\mathscr{L}_g = -\frac{1}{N_r}\sum_{l=1}^{N_r} \log p(r_l \mid r_{<l}, \mathbf{E}), \tag{12}$$

where N_r denotes the length of the target response.

[t] Training Procedure.

Input: training set $\mathscr{P}$for optimizing the model $\mathscr{F}$, hyperparameters $\{\lambda_1, \lambda_2\}$.
Output: Parameters $\mathbf{\Theta}$.

[1] Initialize parameters: $\mathbf{\Theta}$ Initialize memory bank: $\mathscr{M}$ Randomly sample a batch from $\mathscr{P}$. each sample $(\mathscr{D}, R, g, s)$ Add strategy pattern representation into the memory bank $\mathscr{M}$ by Eqn. (6). Update the memory bank $\mathscr{M}$ by Eqn. (7). Update $\mathbf{\Theta}$ by optimizing the loss function in Eqn. (13). $\mathscr{F}$ converges.

4.4. Model Training

Ultimately, we combine all the loss functions for optimizing our model as follows,

$$\mathscr{L} = \mathscr{L}_g + \lambda_1\mathscr{L}_s + \lambda_2\mathscr{L}_r, \tag{13}$$

where $\mathscr{L}$ is the final training objective and λ_1 and λ_2 are the non-negative hyperparameters used for balancing the effect of each loss function on the entire training process. The overall algorithm of the optimization is briefly summarized in the algorithm 4.3.

5. Experiments

5.1. Dataset

We conducted experiments on the ESConv 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. λ_1 and λ_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. Empirically, it takes around 15 epochs to get the peak performance. During the generation stage, we use a maximum of 64 steps to decode the responses. We adopt AdamW [34] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is $2e$-5 and a linear learning rate scheduler with 100 warm-up steps is used to reduce the learning rate progressively. For the framework, we use the PyTorch [35] version 1.8.1 to implement the experiment codes. All experiments were conducted on an NVIDIA Tesla V100 32GB.

5.3. Evaluation Metrics

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

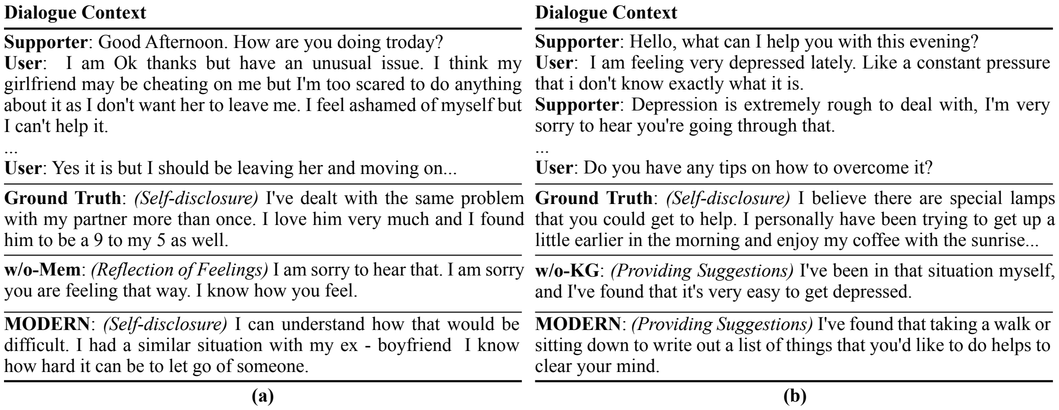
Automatic Evaluation For the automatic evaluation, we adopt several mainstream metrics commonly used in dialogue response tasks, PPL (Perplexity), BLEU-{1,2,3,4} (B-{1,2,3,4}) [36], ROUGE-L (R-L) [37], METEOR (MT) [38], and CIDEr [39].

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

[3] https://huggingface.co/facebook/bart-base.

Table 1
Performance comparison under automatic evaluations. The best results are highlighted in bold.

Model	PPL	B-1	B-2	B-3	B-4	R-L	MT	CIDEr
MoEL	264.11	19.04	6.47	2.91	1.51	15.95	7.96	10.95
MIME	69.28	15.24	5.56	2.64	1.50	16.12	6.43	10.66
DialoGPT-Joint	15.71	17.39	5.59	2.03	1.18	16.93	7.55	11.86
BlenderBot-Joint	16.79	17.62	6.91	2.81	1.66	17.94	7.54	18.04
MISC	16.16	-	7.31	-	2.20	17.91	**11.05**	-
GLHG	15.67	19.66	7.57	3.74	2.13	16.37	-	-
FADO	15.72	-	8.00	4.00	2.32	17.53	-	-
PoKE	15.84	18.41	6.79	3.24	1.78	15.84	-	-
MultiESC	15.41	21.65	9.18	4.99	3.09	20.41	8.84	29.98
MODERN	**14.99**	**23.19**	**10.13**	**5.53**	**3.39**	**20.86**	9.26	**30.08**



Dialogue Context

Supporter: Good Afternoon. How are you doing troday?
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.
...
User: Yes it is but I should be leaving her and moving on...

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.
Supporter: Depression is extremely rough to deal with, I'm very sorry to hear you're going through that.
...
User: Do you have any tips on how to overcome it?

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)

Figure 3: Intuitive comparison of the MODERN and two derivatives.

Table 2
The human evaluation results in four dimensions.

Comparisons	Aspects	Win	Tie	Lose
vs. FADO	Fluency	37.0	24.0	9.0
	Relevance	35.5	20.5	14.0
	Empathy	34.0	16.5	19.5
	Information	37.5	17.0	15.5
vs. MultiESC	Fluency	30.5	24.5	15.0
	Relevance	32.5	20.5	17.0
	Empathy	35.0	19.5	15.5
	Information	37.5	21.0	11.5

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

Table 3
Experimental results of ablation study.

Model	PPL	B-1	B-2	B-3	B-4	R-L	MT	CIDEr
w/o-$\mathscr{L}_s$	15.88	20.25	8.61	4.68	2.91	20.11	8.44	24.32
w/o-$\mathscr{L}_r$	15.32	21.69	9.31	5.06	3.16	20.57	8.87	28.91
w/o-Mem	15.84	21.24	8.90	4.70	2.87	20.21	8.63	27.55
w/o-Emo	15.91	20.35	8.46	4.48	2.72	19.79	8.27	24.01
w/o-KG	15.58	21.56	9.08	4.82	2.96	19.98	8.74	26.83
MODERN	**14.99**	**23.19**	**10.13**	**5.53**	**3.39**	**20.86**	**9.26**	**30.08**

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-$\mathscr{L}_s$**. To show the benefit of constraint for strategy prediction, we removed the corresponding loss function by setting $\lambda_1 = 0$ in Eqn.(13). 2) **w/o-$\mathscr{L}_r$**. 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-$\mathscr{L}_s$ across all metrics. This verifies it is indispensable to constrain the model to predict and respond under proper strategies. 3) w/o-$\mathscr{L}_r$ 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, utilize 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

[1] Z. A. Green, F. Faizi, R. Jalal, Z. Zadran, Emotional support received moderates academic stress and mental well-being in a sample of afghan university students amid covid-19, International Journal of Social Psychiatry 68 (2022) 1748–1755.

[2] C.-W. Chang, F.-p. Chen, Relationships of family emotional support and negative family interactions with the quality of life among chinese people with mental illness and the mediating effect of internalized stigma, Psychiatric Quarterly 92 (2021) 375–387.

[3] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, M. Huang, Towards emotional support dialog systems, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, ACL, 2021, pp. 3469–3483.

[4] W. Peng, Y. Hu, L. Xing, Y. Xie, Y. Sun, Y. Li, Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation, in: Proceedings of the International Joint Conference on Artificial Intelligence, ijcai.org, 2022, pp. 4324–4330.

[5] Y. Cheng, W. Liu, W. Li, J. Wang, R. Zhao, B. Liu, X. Liang, Y. Zheng, Improving multi-turn emotional support dialogue generation with lookahead strategy planning, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACL, 2022, pp. 3014–3026.

[6] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press, 2017, pp. 4444–4451.

[7] Y. Ding, J. Liu, X. Zhang, Z. Yang, Dynamic tracking of state anxiety via multi-modal data and machine learning., Frontiers in psychiatry 13 (2022).

[8] J. Greene, B. Burleson, Handbook of Communication and Social Interaction Skills, American Psychological Association, 2003.

[9] A. C. High, K. R. Steuber, An examination of support (in)adequacy: Types, sources, and consequences of social support among infertile women., Communication Monographs 81 (2014).

[10] B. Wei, S. Lu, L. Mou, H. Zhou, P. Poupart, G. Li, Z. Jin, Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation, in: International Conference on Acoustics, Speech and Signal Processing, IEEE, 2019, pp. 7290–7294.

[11] Y. Liu, W. Bi, J. Gao, X. Liu, J. Yao, S. Shi, Towards less generic responses in neural conversation models: A statistical re-weighting method, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2769–2774. URL: https://aclanthology.org/D18-1297. doi:10.18653/v1/D18-1297.

[12] C. E. Hill, Helping skills: Facilitating, exploration, insight, and action, American Psychological Association, 2009.

[13] C. Zheng, Y. Liu, W. Chen, Y. Leng, M. Huang, Comae: A multi-factor hierarchical framework for empathetic response generation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of *Findings of ACL*, Association for Computational Linguistics, 2021, pp. 813–824.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 7871–7880.

[15] H. Zhou, M. Huang, T. Zhang, X. Zhu, B. Liu, Emotional chatting machine: Emotional conversation generation with internal and external memory, in: Proceedings of the AAAI Conference on Artificial Intelligence, the innovative Applications of Artificial Intelligence, and the AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI Press, 2018, pp. 730–739.

[16] H. Rashkin, E. M. Smith, M. Li, Y. Boureau, Towards empathetic open-domain conversation models: A new benchmark and dataset, in: Proceedings of the Conference of the Association for Computational Linguistics, ACL, 2019, pp. 5370–5381.

[17] L. Shen, Y. Feng, CDL: curriculum dual learning for emotion-controllable response generation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 556–566.

[18] M. Y. Chen, S. Li, Y. Yang, Emphi: Generating empathetic responses with human-like intents, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, 2022, pp. 1063–1074.

[19] W. Kim, Y. Ahn, D. Kim, K. Lee, Emp-rft: Empathetic response generation via recognizing feature transitions between utterances, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, 2022, pp. 4118–4128.

[20] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, Z. Chen, Empdg: Multi-resolution interactive empathetic dialogue generation, in: Proceedings of the International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2020, pp. 4454–4466.

[21] N. Majumder, P. Hong, S. Peng, J. Lu, D. Ghosal, A. F. Gelbukh, R. Mihalcea, S. Poria, MIME: mimicking emotions for empathetic response generation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACL, 2020, pp. 8968–8979.

[22] Z. Lin, A. Madotto, J. Shin, P. Xu, P. Fung, Moel: Mixture of empathetic listeners, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing, ACL, 2019, pp. 121–132.

[23] X. Xu, X. Meng, Y. Wang, Poke: Prior knowledge enhanced emotional support conversation with latent variable, CoRR abs/2210.12640 (2022).

[24] Q. Tu, Y. Li, J. Cui, B. Wang, J. Wen, R. Yan, MISC: A mixed strategy-aware model integrating COMET for emotional support conversation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL, 2022, pp. 308–319.

[25] W. Peng, Z. Qin, Y. Hu, Y. Xie, Y. Li, FADO: feedback-aware double controlling network for emotional support conversation, Knowledge-Based Systems 264 (2023) 110340.

[26] W. J. Reynolds, B. Scott, Empathy: a crucial component of the helping relationship, Journal of psychiatric and mental health nursing 6 (1999) 363–370.

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

Social Networking 21 (2018) 625–636.

[28] T. Parkin, A. de Looy, P. Farrand, Greater professional empathy leads to higher agreement about decisions made in the consultation, Patient Education and Counseling 96 (2014) 144–150.

[29] Q. Li, P. Li, Z. Ren, P. Ren, Z. Chen, Knowledge bridging for empathetic dialogue generation, in: The Conference on Artificial Intelligence, Conference on Innovative Applications of Artificial Intelligence, The Symposium on Educational Advances in Artificial Intelligence, AAAI Press, 2022, pp. 10993–11001.

[30] Y. Liu, W. Maier, W. Minker, S. Ultes, Empathetic dialogue generation with pre-trained roberta-gpt2 and external knowledge, in: Conversational AI for Natural Human-Centric Interaction International Workshop on Spoken Dialogue System Technology, volume 943 of *Lecture Notes in Electrical Engineering*, Springer, 2021, pp. 67–81.

[31] L. Jing, X. Song, X. Lin, Z. Zhao, W. Zhou, L. Nie, Stylized data-to-text generation: A case study in the e-commerce domain, ACM Trans. Inf. Syst. (2023).

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 2017, pp. 5998–6008.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Brew, Huggingface's transformers: State-of-the-art natural language processing, CoRR abs/1910.03771 (2019).

[34] I. Loshchilov, F. Hutter, Fixing weight decay regularization in adam, CoRR abs/1711.05101 (2017).

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems, 2019, pp. 8024–8035.

[36] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL, 2002, pp. 311–318.

[37] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[38] A. Lavie, A. Agarwal, METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 228–231.

[39] R. Vedantam, C. L. Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Conference on Computer Vision and Pattern Recognition, IEEE, 2015, pp. 4566–4575.

[40] N. Schluter, The limits of automatic summarisation according to ROUGE, in: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, ACL, 2017, pp. 41–45.

[41] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, B. Dolan, DIALOGPT : Large-scale generative pre-training for conversational response generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, 2020, pp. 270–278.

[42] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, J. Weston, 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, Association for Computational Linguistics, 2021, pp. 300–325.

[43] K. Sohn, H. Lee, X. Yan, Learning structured output representation using deep conditional generative models, in: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems, 2015, pp. 3483–3491.

[44] S. M. Mohammad, Obtaining reliable human ratings of valence, arousal, and dominance for 20, 000 english words, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL, 2018, pp. 174–184.