electronics

Article

AdMISC: Advanced Multi-Task Learning and Feature-Fusion for Emotional Support Conversation

Xuhui Jia [1], Jia He [1,*], Qian Zhang [2] and Jin Jin [3]

[1] School of Computer Science, Chengdu University of Information and Technology, Chengdu 610225, China; jxh@xugudb.com
[2] Active Network (Chengdu), Ltd., Chengdu 610021, China; kernel.zhang@activenetwork.com
[3] School of Software Engineering, Chengdu University of Information and Technology, Chengdu 610225, China; jinjin@cuit.edu.cn
* Correspondence: hejia@cuit.edu.cn

Abstract: The emotional support dialogue system is an emerging and challenging task in natural language processing to alleviate people's emotional distress. Each utterance in the dialogue has features such as emotion, intent, and commonsense knowledge. Previous research has indicated subpar performance in strategy prediction accuracy and response generation quality due to overlooking certain underlying factors. To address these issues, we propose Advanced Multi-Task Learning and Feature-Fusion for Emotional Support Conversation (AdMISC), which extracts various potential factors influencing dialogue through neural networks, thereby improving the accuracy of strategy prediction and the quality of generated responses. Specifically, we extract features affecting dialogue through dynamic emotion extraction and commonsense enhancement and then model strategy prediction. Additionally, the model learns these features through attention networks to generate higher quality responses. Furthermore, we introduce a method for automatically averaging loss function weights to improve the model's performance. Experimental results using the emotional support conversation dataset ESConv demonstrate that our proposed model outperforms baseline methods in both strategy label prediction accuracy and a range of automatic and human evaluation metrics.

Keywords: multi-task learning; dialog generation; emotional support conversation; attention

check for updates

Citation: Jia, X.; He, J.; Zhang, Q.; Jin, J. AdMISC: Advanced Multi-Task Learning and Feature-Fusion for Emotional Support Conversation. *Electronics* **2024**, *13*, 1484. https://doi.org/10.3390/electronics13081484

Academic Editor: Arkaitz Zubiaga

Received: 11 March 2024
Revised: 6 April 2024
Accepted: 8 April 2024
Published: 13 April 2024

1. Introduction

In the ever-evolving landscape of society, individuals are encountering increasing mental stress in their daily lives. The research shows that over 50% of adults have grappled with mental illnesses or disorders at some point, yet only approximately 20% of these individuals have sought or received relevant treatment. Recent studies have highlighted the growing significance of emotional support conversation (ESC) as a form of mental health therapy, garnering considerable attention [1]. More and more researchers are integrating emotional support conversation with dialogue systems as a novel, intelligent mental-health therapy approach, and it has been applied to fields such as intelligent customer service and intelligent psychological counseling, such as Woebot [2]. This emerging field paves the way for innovative developments in dialogue systems and offers a promising avenue for addressing mental health challenges.

As shown in Figure 1, the emotional support conversation takes place through multiple dialogue rounds between the seeker and the supporter. It requires supporters to employ a specific support strategy to respond empathetically to alleviate the seeker's distress. The existing research mainly focuses on two aspects: firstly, the accurate prediction of dialogue strategies to tailor responses accordingly. For instance, Tu et al. [3] utilized a mixed strategy prediction method. Secondly, the enhancement of the model's comprehension of dialogue context, such as the work of Peng et al. [4] who designed a hierarchical graph network to capture user intent.

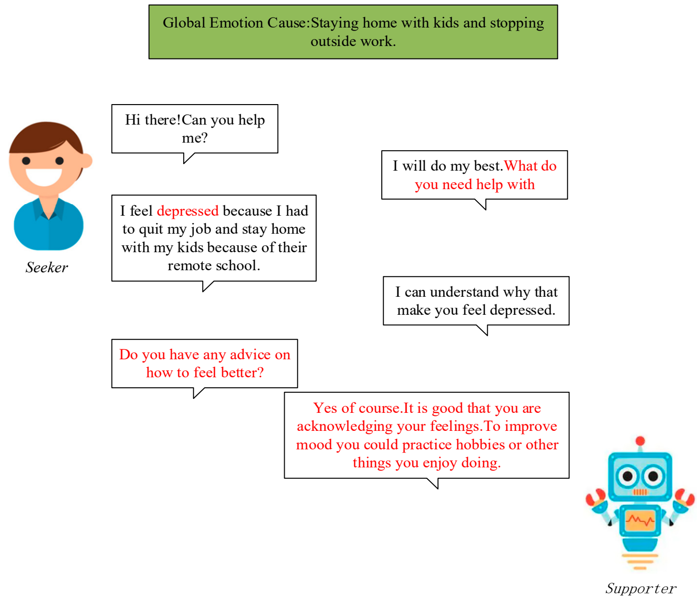
Global Emotion Cause:Staying home with kids and stopping outside work.
Hi there!Can you help me?
I will do my best.What do you need help with
I feel depressed because I had to quit my job and stay home with my kids because of their remote school.
I can understand why that make you feel depressed.
Do you have any advice on how to feel better?
Yes of course.It is good that you are acknowledging your feelings.To improve mood you could practice hobbies or other things you enjoy doing.
Seeker
Supporter

Figure 1. An example of an emotional support conversation from the ESConv dataset. The red font in the figure indicates the characteristics of emotional support conversation, the supporter expresses its willingness to help, the seeker explains its emotional state and emotional support needs, supporter provides comfort and advice.

Despite some of the achievements of researchers, the task still faces the following challenges:

1. As the conversation progresses, users' emotions subtly evolve. Accurately identifying these emotional changes is essential for the model to predict strategy labels and provide empathetic responses [5].
2. Dialogue strategy as a linguistic pattern is a highly complex concept encompassing many language features [6]. Previous studies have modeled it using a single vector (i.e., category labels), which is insufficient for fully representing the complexity of strategy information. Integrating the contextual information that influences strategies has become a challenge.
3. Existing emotional support dialogue models tend to generate generalized responses [7], which fail to provide effective emotional support. To address this issue, introducing more contextually relevant concepts can facilitate the model in generating more meaningful suggestions tailored to specific situations. How to explore these relevant concepts and integrate them is a crucial task.
4. In multi-task joint training, the model's performance heavily relies on the weights assigned to each task's loss function [8], posing challenges in manual weight adjustment.

In response to the issues above in the current research, we propose a method of multi-task learning and feature-fusion in emotional support conversations, termed AdMISC. This method is based on the pre-trained transformer neural network model [9], and addresses the identified problems. The main contributions of our work are as follows:

1. Addressing the oversight of dynamic emotional changes in existing models, AdMISC incorporates an Emotion Detector module to detect these changes. This utilization of dynamic emotional characteristics guides strategy prediction learning effectively.
2. To address the limitations of single-vector modeling in strategy prediction, we propose a mixed strategy approach, which utilizes neural networks to enhance dialogue history and problem descriptions with commonsense reasoning. Additionally, it integrates commonsense-enhanced information and dynamic emotional information to jointly model strategy prediction.
3. To alleviate the generalization tendency observed in the generated text of existing models during the emotional support generation stage, we propose a feature fu-

sion approach. This method leverages neural network multi-head attention and cross-attention mechanisms to focus on the original dialogue history, commonsense-enhanced dialogue history, commonsense-enhanced problem descriptions, dynamic emotional information, and strategy selection information in the feedforward network. These context-related concepts can guide the model in generating more targeted and suggestive responses.

4. We propose a dynamic multi-task loss function weight balancing method to address the challenge of manually adjusting task weights in multi-task joint training. This method balances the impact of multiple loss functions on model training.

The experimental results demonstrate that the AdMISC model outperforms other baseline models in both automatic and human evaluation metrics on the ESConv dataset, confirming the feasibility and effectiveness of our approach.

2. Related Work

2.1. Conversation Strategy

In emotional support dialogue systems, the selection of dialogue strategies plays a pivotal role in shaping the seeker experience, as distinct strategies yield varied response generation outcomes [10]. Existing emotional support dialogue systems commonly utilize deep learning for strategy selection. For instance, Tu et al. [3] proposed a mixed strategy learning method grounded in deep learning principles. On the other hand, Peng et al. [11] incorporated seeker emotional feedback information for dialogue strategy selection. Xu et al. [12] employed a prior knowledge method in predicting dialogue strategy labels, and Cheng et al. [13] considered forward-looking heuristic strategy planning and selection.

Despite their remarkable achievements, the existing work still faces the challenges of intricate strategy modeling and variability of emotions. Furthermore, Zeng et al. [14] noted that strategy selection is intricately linked to context. Integrating implicit information present in the conversational context becomes imperative when classifying strategies.

2.2. Emotional Response Generation

The emotional response generation module within the emotional support dialogue system produces responses imbued with emotional support meaning, aligned with the selected dialogue strategy, and delivers them back to the seeker. Recent studies have suggested that augmenting the generation process with additional information can enhance the overall performance of emotional response generation. For instance, Zhong et al. [15] leveraged the ConceptNet [16] module to enhance response generation and emotional states. Quan et al. [3] captured the seekers' mental state by incorporating a generative commonsense model COMET [17], interacting with various factors to generate emotional responses. Deng et al. [18] enhanced the system through knowledge in the field of mental health. Other studies focus on acquiring the seeker's situations, emotions, and intentional information. For example, Xu et al. [12] explored contextual semantic relations and emotional states, while Zhao et al. [19] considered the transformation of semantics, strategies, and emotions in the model.

Although the improvement of the above method allows the model to generate fluent text and significantly reduces the occurrence of logical errors, there are still challenges in the emotional support dialogue task: replies tend to be general and lack pertinence and suggestions for questions. It is difficult to achieve the purpose of emotional support. In this regard, Wang et al.'s research [20] emphasized that the logic of emotional dialogue replies should prioritize improving references in context and strengthening the connections among themes, emotions, and knowledge. This approach aims to generate replies that align more closely with thematic logic, offer accurate references, and convey rich emotion. Additionally, Wang et al. [21] proposed that the language model should iteratively infer the psychological and emotional state information of the interlocutor based on the dialogue history as the thinking chain, thereby enhancing the quality of responses.

2.3. Influencing Features in ESC

In the research of emotional support conversations, the current work largely relies on seekers' emotional labels and contextual cues to guide models in perceiving emotional information within dialogues. However, due to the multifaceted nature of human emotion perception and expression, models trained solely on emotional labels and contextual cues may overlook these underlying influential factors. According to Yang et al. [22], a range of potential dialogue information could impact the effectiveness of models in learning strategy selection and generating emotionally supportive responses. These factors include conversation-level emotions, sentence-level emotions, seeker intentions in inquiries, dialogue history, and human common sense involved in the conversation. Additionally, psychological studies by Hill et al. [23] indicated that emotional support conversations involve a complex interactive process requiring consideration of various information, such as seeker emotions, intentions, emotional fluctuations, and commonsense content, which can be mined from dialogue contexts.

2.4. Commonsense Knowledge Generation Model COMET

To enhance the model's understanding of emotional support conversations using additional knowledge, past approaches have typically utilized pre-constructed commonsense knowledge bases or semantic networks, applying known relationships from these knowledge bases to entities within the dialogue. However, Bosselut et al. [17] argued that commonsense knowledge does not entirely suit the pattern of combining two entities with known relationships, and instead, they proposed using an automatically constructed knowledge base to generate commonsense knowledge. Therefore, they introduced a commonsense knowledge generation model called COMET, based on a large-scale pre-trained transformer neural network. This model can adaptively generate knowledge representations, meaning that given a head entity s and a tail entity o, it generates a relation r, forming high-quality commonsense semantic relation triples $\{s, r, o\}$. These relations are derived from the sets of relations defined in ConceptNet. Once trained, the COMET model can generate reasonable, rich, and novel commonsense semantic triples, even when faced with commonsense events unseen by the model.

3. Task Definition

In the training of the emotional support dialogue model, considering our training dataset, it can be articulated as follows:

$$D = \{p_1, p_2, \dots p_M\} \tag{1}$$

composed of M samples. Each sample is composed as follows:

$$p_i = \{S_i, C_i, h_i, R_i\} \tag{2}$$

including S_i as the seeker's situation; C_i as a dialogue context; h_i as a strategy for supporting; and R_i as a target response. C_i contains the history utterances between seeker and supporter, R_i and C_i as follows:

$$C_i = (CLS, u_1^i, EOS, u_2^i \dots u_N^i) \tag{3}$$

$$R_i = (r_1^i, r_2^i, \dots, r_N^i) \tag{4}$$

where CLS is the start-token and also describes the state. EOS is the separation token between two utterances. C_i and R_i include N tokens. The goal of the ESC task is to build a model F that can generate an expected supportive response r_g^i referring the C_i and S_i as:

$$r_g^i = F(C_i, S_i \middle| \Theta) \tag{5}$$

where Θ is the set of learned parameters of F.

4. Method

The comprehensive architecture of AdMISC is depicted in Figure 2. The process begins with obtaining information about the conversation through the encoder. At this stage, each sentence underwent emotion recognition via the Emotion Detector. The natural language labels corresponding to each emotion were then amalgamated to derive dynamic emotional changes. Simultaneously, COMET processed the seeker's situation and the seeker's last reply for commonsense enhancement. This additional information and the dialogue context served as inputs to the encoder.

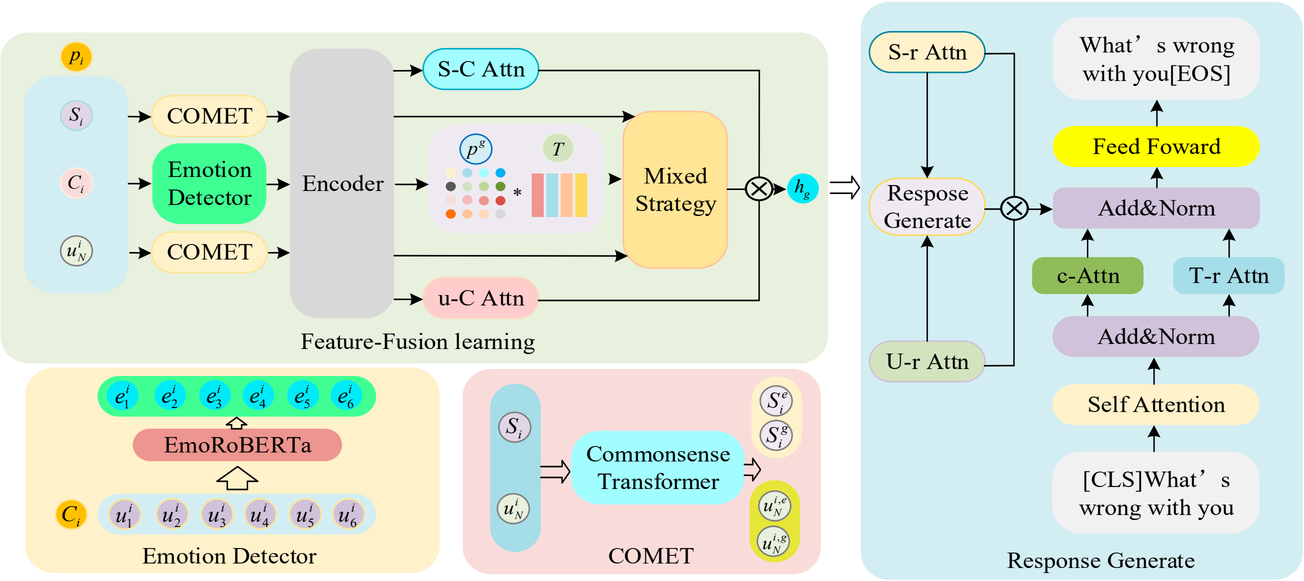

Figure 2. The overall architecture of our proposed AdMISC model mainly consists of two modules: Feature-Fusion learning and Response Generate. The Feature-Fusion learning module also contains two sub-modules: Emotion Detector with EmoRoBERTa to recognize the specific emotion in the seeker's utterances, and COMET to generate commonsense knowledge based on the conversation. * in the figure represents linear matrix multiplication.

Subsequently, within the Mixed-Strategy learning module, fine-grained dynamic emotional information, commonsense-enhanced historical conversations, and commonsense-enhanced conversation descriptions were integrated to model strategy prediction. Finally, the information obtained above was focused through a multi-layer attention network and injected into the emotional expression generation module and Decoder.

4.1. Emotion Detector

Mental health research underscores the significance of empathy in emotional support [24] and emphasizes that an essential aspect of enhancing empathy capabilities involves providing fine-grained emotional information [25]. Consequently, in training emotional support dialogue systems, it proves highly advantageous for the model to gain a coherent understanding of the seeker's emotional state by capturing dynamic and fine-grained emotional changes, as opposed to relying solely on static emotional signals. To address this, we proposed the Emotion Detector module to discern the dynamic changes in the seekers' fine-grained emotions throughout the conversation.

Specifically, we utilized a BERT-based pre-trained emotion detection model, EmoRoBERTa [26], capable of discerning the emotion categories present in the input text. The model's output comprises 28 distinct emotions; we integrated these emotions to correspond with the 7 emotions in the dataset. Emotion recognition was executed by inputting each utterance within the ongoing round of dialogue context text into EmoRoBERTa, as expressed by the following equation:

$$e_j = E(u_j) \tag{6}$$

where the predicted emotion category word from the model is employed to signify the emotion detected in a conversation; these emotional category words are subsequently input into the encoder in their natural language form. This methodology circumvented the introduction of unnecessary parameters that could potentially disrupt model learning and is articulated as follows:

$$E_j = (e_1, e_2, \ldots, e_N) \tag{7}$$

The emotional support dialogue model can effectively extract the dynamic emotional changes corresponding to the dialogue process by employing the method above.

4.2. Commonsense Enhance

We utilized COMET to generate commonsense knowledge for S_i and u_N^i, a process that can be represented as follows:

$$[S_i^g, S_i^e] = COMET(S_i) \tag{8}$$

$$[u_N^{i,g}, u_N^{i,e}] = COMET(u_N^i) \tag{9}$$

where S_i^e and $u_N^{i,e}$ represent the parts of the generated common sense with emotional factors; S_i^g and $u_N^{i,g}$ represents the remaining part. Afterwards, we inputted them with C_i into the attention network, S-C Attn and u-C Attn, allowing the model to learn the commonsense knowledge part, which can be represented as:

$$A^s = CROSS - ATT([S_i^g, S_i^e], C_i) \tag{10}$$

$$A^x = CROSS - ATT([u_N^{i,g}, u_N^{i,e}], C_i) \tag{11}$$

After obtaining the representation enhanced with commonsense knowledge, we inputted CLS into a multi-layer perceptron MLP to obtain a probability distribution p^g for representing the strategy. Multiplying p^g with the strategy labels T from the dataset, we can obtain the initial strategy selection h_g. The process is described as follows:

$$p^g = MLP(CLS) \tag{12}$$

$$h_g = p^g T \tag{13}$$

Through the steps above, the model obtained commonsense knowledge from the dialogue via COMET and the initial predictions of strategy labels.

4.3. Feature-Fusion Learning to Predict Strategy Labels

Existing emotional support dialogue models commonly employ a single vector (i.e., seeker's emotional state) modeling method during strategy selection learning. However, the process of dialogue strategy learning is a multifaceted concept encompassing various language features [7]. The modular approach of a single vector model proves insufficient for adequately representing intricate strategy pattern information. We proposed the Mixed-Strategy module to capture information in the dialogue and effectively model strategy selection.

This module integrated an initial strategy representation with dialogue history, commonsense-enhanced seeker's last reply, emotion cause composed of commonsense-enhanced dialogue descriptions, and dynamic sentence-level emotional state, thus modeling the strategy prediction learning process by combining the above information. The network structure of the module is depicted in Figure 3.

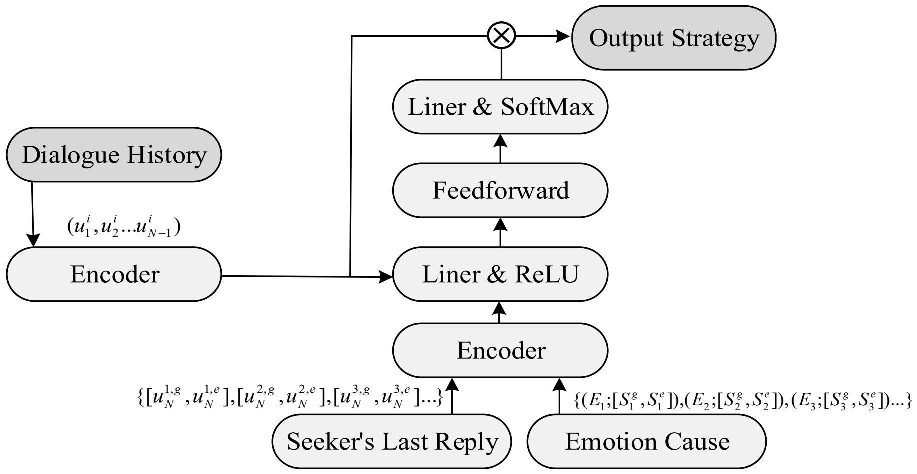
Output Strategy
Liner & SoftMax
Feedforward
Dialogue History
$(u_1^i, u_2^i \ldots u_{N-1}^i)$
Encoder
Liner & ReLU
Encoder
$\{[u_N^{1,g}, u_N^{1,e}], [u_N^{2,g}, u_N^{2,e}], [u_N^{3,g}, u_N^{3,e}] \ldots\}$
$\{(E_1;[S_1^g, S_1^e]), (E_2;[S_2^g, S_2^e]), (E_3;[S_3^g, S_3^e]) \ldots\}$
Seeker's Last Reply
Emotion Cause

Figure 3. Network structure of Mixed-Strategy module. The dialogue history, coupled with the emotion cause and the seeker's last reply enriched by common sense, serve as inputs to the encoder. Following processing through various network layers, the predicted strategy labels are derived.

The dialogue history information undergoes embedding through the transformer encoder to obtain the corresponding representation, which is expressed as:

$$H_t = \left\{ C_i; [u_N^{i,g}, u_N^{i,e}] \right\}_{(i=1)}^{(t-1)} \tag{14}$$

Similarly, linearly combining the seeker's situation of each round of dialogue with the sentence-level dynamic emotional states obtained through the Emotion Detector, we can obtain the sequence of Emotion Cause U_t, which is expressed as:

$$U_t = \left\{ E_i; [S_i^g, S_i^e] \right\}_{(i=1)}^{(t-1)} \tag{15}$$

The above results were combined into a long vector in the Contact layer and input into a linear layer. The result was obtained through the ReLU activation function, which can be expressed as:

$$\mu = ReLU(W_\mu [H_t; U_t] + b_\mu) \tag{16}$$

among them, W_μ and b_μ are trainable parameters. After μ is obtained, perform a weighted operation on them and the comprehensive strategy representation, expressed as:

$$\hat{h}_g = \mu \cdot H_t + (1 - \mu) \cdot U_t \tag{17}$$

$\hat{h}_g$ in the feed-forward network with residual connections in the input sublayer, the hidden state generated is expressed as $\tilde{h}_g$, and it is used as input. The output is obtained $\overset{\smile}{h}_g$ through the SoftMax activation function, which can be described as:

$$\tilde{h}_g = \sigma(\hat{h}_g) \tag{18}$$

$$\overset{\smile}{h}_g = softmax(W_s \tilde{h}_g + b_s) \tag{19}$$

among them, σ represents the hidden layer calculation, and is a trainable parameter. The obtained information was multiplied with the comprehensive strategy representation $\overset{\smile}{h}_g$, updating the strategy h_g, expressed as:

$$h_g = \beta \cdot \overset{\smile}{h}_g + h_g \tag{20}$$

β is a hyperparameter.

To train the obtained improved comprehensive strategy, the negative log-likelihood estimate of the ground truth real strategy label was used as its loss function, expressed as:

$$L_g = -log(p(h_i \Big| h_g, C_i, S_i, u_N^i)) \tag{21}$$

This module used the encoder structure in the transformer network to encode the dialogue history and emotion cause information in the dialogue text and combined it with the initial strategy label to obtain a new strategy label.

4.4. Fusion of Dialogue Features to Generate Responses

To mitigate the generalization tendency of emotionally supportive response texts generated by the model, we proposed a feature-fusion-based response generation method. Specifically, we utilized commonsense-enhanced seeker descriptions, the seeker's last reply, dynamic emotional states obtained from the Emotion Detector, strategy labels obtained through the Mixed-Strategy network, and the dialogue history to guide the model's response generation via attention mechanisms. The emotional responses generation module Response-Generate was proposed. Its network structure is shown in Figure 4.

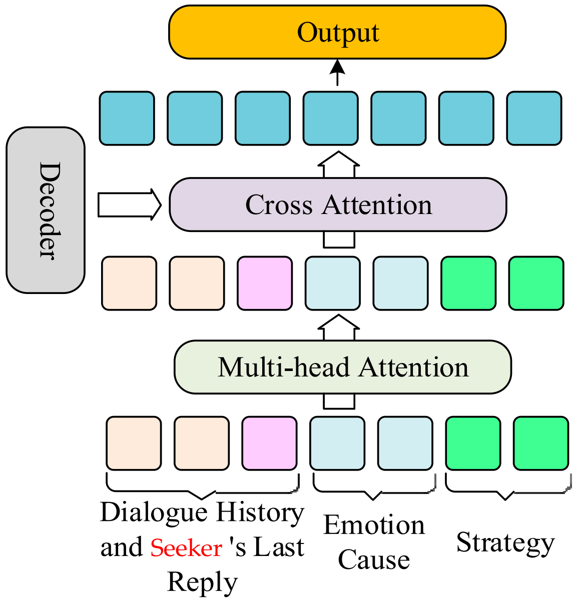

Figure 4. The network structure of Response-Generate contains three parts: the commonsense-enhanced dialogue history and seeker's last reply, strategy, and emotion cause as input. Combine them with different layers, then update the hidden state of the decoder. Finally, the output is a generated response with feature fusion.

Input H_t and h_g, U_t and h_g into the multi-head attention layer for attention enhancement, which can be expressed as:

$$\breve{H}_t = MH - ATT(H_t, h_g) \tag{22}$$

$$\breve{U}_t = MH - ATT(U_t, h_g) \tag{23}$$

combined $\breve{H}_t$ and $\breve{U}_t$ with the hidden state O of the decoder through the cross-attention network, specifically as follows:

$$A^h = CROSS - ATT(\breve{H}_t, O) \tag{24}$$

$$A^u = CROSS - ATT(\overset{\smile}{U}_t, O) \tag{25}$$

Similarly, we enhanced O with attention to improve the model's learning ability of commonsense knowledge during response generation. The representation is as follows:

$$A^r = CROSS - ATT([S_i^g, S_i^e], O) \tag{26}$$

$$A^c = CROSS - ATT([u_N^{i,g}, u_N^{i,e}], O) \tag{27}$$

The updated label of strategy h_g obtained was inputted along with O into the cross-attention network, which can be represented as:

$$A^g = CROSS - ATT(h_g, O) \tag{28}$$

The new information obtained was combined with other information of the model, expressed as:

$$O' = LN(O + A^g + A^x + A^s + A^c + A^r + A^u + A^h) \tag{29}$$

We combined all the information enhanced through attention mechanisms with O, resulting in a fused representation O' that incorporated various dialogue information. This fused representation guided the model in generating the final response.

Similarly, the ground truth target reply used negative log-likelihood estimation as the loss function for training the final reply, which can be expressed as:

$$L_r = -\sum_{t=1}^{n_r} log(p(r_t | r_{j<t}, C_i, S_i, u_N^i)) \tag{30}$$

n_r is the length of the reply.

This module adopted the decoder structure, multi-head attention mechanism, and cross-attention mechanism. It integrated crucial dialogue information to enhance the quality of the response generated by the model.

4.5. Multi-Task Joint Training Loss Function

In existing multi-task jointly trained neural-network emotional support dialogue system models, there are primarily two approaches to handling the loss function:

1. Determine the relationship between each loss function during the initial training by manually tuning the weights, which remain constant throughout the training process.
2. Observe changes in various indicators of the loss function during the network training process and manually adjust the weights accordingly.

However, the performance of multi-task joint learning models is highly dependent on these weights, making the process of finding optimal weights through manual adjustment complex and challenging. Building on the proposition of Peng et al. [27] that the loss function of multi-task joint training should be assigned specific weights, this paper introduced a hyperparameter for each loss function of the two tasks, assigning them distinct weights. This optimization aimed to enhance the model's performance, specifically:

$$L = \alpha_1 L_r + \alpha_2 L_g \tag{31}$$

α_1 and α_2 are the weight of the two loss functions. Treat the minimization of L as the optimization goal of the entire model.

To dynamically adjust the weight values, inspired by the research conducted by Liu [28] and others, this paper introduced a dynamic weight averaging method. This method involves learning the average by considering each task's loss-changing rate. The methodology can be expressed as follows:

$$\alpha_k(t) = \frac{Ke^{(w_k(t-1)/T)}}{\sum_i e^{(w_i(t-1)/T)}} \tag{32}$$

$$w_k(t-1) = \frac{L_k(t-1)}{L_k(t-2)} \tag{33}$$

$w_k(\cdot)$ is the relative decline rate; t is the iteration index; T represents the tasks used to control temperature. T will make the distribution between different tasks more even. $L_k(t)$ is the average loss of each epoch in the iteration step. On the initial training set, first initialize $w_k(t)$ to 1 based on experience. After obtaining the final weight result, evaluate it on the verification set to select the optimal weight.

5. Experiments

5.1. Experimental Setup

Platform Settings. We employed the proposed AdMISC model to conduct experiments on the emotional support dialogue dataset ESConv and perform ablation experiments on the above modules. The details of the experimental environment are outlined in Table 1.

Table 1. Experimental platform settings.

Environments	Details
OS	Ubuntu 16.04 LTS 64 bit
CPU	Intel(R) Xeon(R) Silver 4210
GPU	NVIDIA TITAN RTX
Memory	125 GB
Platform	Python 3.8, PyTorch

Dataset and processing. We conducted experiments on the ESConv [29] dataset, which is a high-quality dataset for emotional support conversation tasks. This is an English dataset. The builder recruited multiple crowd-workers who understood emotional support conversation procedures and strategies to talk to volunteers with emotional support needs through an online platform. The crowd-workers conducted interviews on the strategies adopted in each round of conversation. Annotations were made, and seekers provided feedback on their emotional status after every two conversation rounds, indicating reduced emotional distress. Each sample in the dataset is a conversation between a seeker and a supporter, and each conversation contains additional information, such as a description of the problem faced by the seeker and annotations of strategy categories in the supporter's response. The conversation unfolded in three stages: inquiry, reassurance, advice, and finally, an assessment of the intensity of the seeker's current emotions. The dataset contains 1300 long conversations, with an average of 29.5 utterances per conversation. The conversations have a total of 5 topics and 7 emotions, as well as 8 support strategies. To facilitate a more effective comparison with the baseline model, this paper performed similar preprocessing on ESConv. The conversation samples were truncated every 10 rounds and randomly divided into training, validation, and test sets in a ratio of 8:1:1. The detailed statistics are shown in Table 2.

Implementation Details. We implemented the specific process based on blenderbot-small [30], utilizing AdamW as the optimizer, adjusted parameters, and incorporated the Dropout mechanism to mitigate overfitting. The detailed experimental parameter settings are presented in Table 3.

Table 2. Processed to ESConv Dataset.

Category	Training Set	Valid Set	Test Set
Total of sentence	14,117	1764	1764
Average number of words per sentence	17.25	17.09	17.11
Number of tie sentences in a single round of dialogue	7.61	7.58	7.49
The average number of words in a single conversation	148.46	146.66	145.17

Table 3. Experimental parameter settings.

Parameters	Values
block size	512
batch_size	20
adam_epsilon	0.000000001
epoch	8
Dropout	0.1
learning_rate	0.00002

5.2. Evaluation Metrics

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

Automatic Evaluation. We employed a set of automatic evaluation indicators to assess the performance of the proposed model and other baseline models. These included:

1. Strategy Prediction Accuracy (ACC): This metric evaluates the model's accuracy in strategy prediction. For the same dataset, a higher ACC indicates more accurate predictions by the model.
2. Perplexity (PPL): Perplexity measures how well the model predicts the sequence of words. A lower perplexity score indicates better performance.
3. BLEU-2 (B-2) and BLEU-4 (B-4): These scores represent the similarity between the response generated by the model and the ground truth. Higher BLEU scores indicate better alignment with the real answer.
4. ROUGE-L (R-L): This metric evaluates the overlap of words and sequences between the generated response and the ground truth. A higher ROUGE-L score signifies a closer resemblance to the real answer.

Human Evaluation. In order to comprehensively evaluate the improvement effect of the AdMISC emotional support dialogue system, this study employed human evaluation involving real participants. The evaluation method included engaging 5 participants to assume the role of seeker and interact with the AdMISC, FADO, and MISC models. Participants assessed the performance of the two models in specific scenarios, with agreement required from at least half of the participants before counting. An additional reviewer was invited to conduct random sampling and review 10% of the assessment results to ensure assessment quality. Specific aspects of the assessment included:

1. Fluency: determining which model can generate more fluent and coherent responses.
2. Accuracy: assessing which model better identifies the seeker's problem.
3. Empathy: evaluating which model better understands the seeker's feelings and situation.
4. Suggestion: analyzing which model provides more effective suggestions.
5. Overall: considering which model provides a more effective emotional support effect.

5.3. Baselines

This is a brief introduction to the AdMISC model proposed in this article and other comparative models. The parameters are all default settings.

1. Transformer [9]: is a common Seq2Seq model trained based on the MLE loss function;

2. MT Transformer [31]: this takes sentiment prediction as an additional learning task and uses sentiment labels provided in ESConv to learn sentiment prediction;
3. MoEL [32]: combining output states from multiple decoders to enhance empathetic reply generation for different emotions;
4. MIME [33]: considers polarity-based emotion clustering and emotion mimicry to generate empathic replies;
5. BlenderBot-Joint [30]: preset a special strategy token before generating an emotional support reply statement;
6. MISC [3]: emotionally supportive dialogue model based on ESConv predicts emotional labels and generates emotionally supportive replies through a hybrid strategy learning module;
7. GLGH [4]: this model establishes a global-to-local hierarchical graph structure to generate supportive emotional responses through the seeker's global emotional states and local intentions;
8. FADO [11]: this model designs a two-level feedback strategy selector to punish or encourage the strategy during the strategy selection process.

5.4. Model Comparison and Analysis

We conducted comparative experiments between AdMISC and baseline models from automatic and human evaluation perspectives.

Automatic Evaluation. The experimental results comparing the AdMISC model with the above baseline models are presented in Table 4. The results indicate the following:

1. AdMISC outperforms the baselines in most metrics, which is powerful proof of the effectiveness of the proposed method.
2. Models that come after BlenderBot-Joint combine the dialogue history with static emotion labels to guide strategy prediction, resulting in improved performance on ACC compared to previous models. AdMISC, in addition to the information above, incorporates additional dialogue information, further enhancing the ACC metric. This demonstrates that our approach of mining additional information and integrating them into the model is effective.
3. In the remaining metrics related to dialogue diversity and fluency, AdMISC also excels, indicating that the response generation module still requires strategy and other information from the dialogue, such as commonsense knowledge and emotion, to facilitate the generation of supportive responses further.

Table 4. Comparison between AdMISC and baseline models. The upward arrow in the figure indicates that the higher the evaluation standard, the better, and the downward arrow indicates that the lower the evaluation standard, the better.

Models	ACC (%)	PPL ↓	D-1 ↑	B-2 ↑	B-4 ↑	R-L ↑
Transformer	-	89.61	1.29	6.53	1.37	15.17
MT Transformer	-	89.52	1.28	6.58	1.47	14.75
MoEL	-	133.13	2.33	5.93	1.22	14.65
MIME	-	47.51	2.11	5.23	1.17	14.74
BlenderBot-Joint	28.57	18.49	3.12	5.78	1.74	16.39
MISC	31.01	16.16	3.41	6.24	1.76	17.37
GLGH	-	15.72	3.50	7.57	2.13	16.37
FADO (SOTA)	32.90	15.52	3.80	8.00	2.32	17.53
AdMISC	**34.41**	**15.49**	**3.83**	**8.16**	**2.36**	**17.91**

Human Evaluation. The evaluation results, expressed as the percentage of participants choosing a certain model out of the total, are presented in Table 5. We report the comparison results between our model and the two baselines (i.e., FADO and MISC). In particular, for each pair of model comparisons and each metric, we show the number of samples where our model achieves a better (denoted as "Win"), equal (denoted as "Tie"), and worse performance (denoted as "Lose") compared with the baselines. As seen, AdMISC 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 4. In addition, the number of "Win" cases is the largest for the Suggestion metric compared with other metrics, which demonstrates that integrating all the methods we proposed can supply meaningful information for emotional support.

Table 5. Human evaluation results.

AdMISC vs.	FADO (SOTA)			MISC		
	Win	**Lose**	**Tie**	**Win**	**Lose**	**Tie**
Fluency	**23.6**	20.7	55.7	**44.7**	18.3	37.0
Accuracy	**26.0**	25.3	48.7	**37.0**	16.3	46.7
Empathy	**27.7**	26.3	46.0	**53.7**	7.0	39.3
Suggestion	**28.7**	25.0	46.3	**37.7**	27.3	35.0
Overall	**22.3**	17.5	60.2	**64.7**	17.0	18.3

5.5. Ablation Study

We compared the original AdMISC model with the following derived model and proved that all designed modules played a certain role by comparing the changes in ACC, D-1, B-2, R-L, Precision, and Recall.

w/o *u*. To show the benefit of the Mixed-Strategy learning module, we removed the corresponding loss function by setting $\alpha_1 = 0$ in Equation (31).

w/o *a*. To show the effect of the Response-Generate module, we removed the corresponding loss function by setting $\alpha_2 = 0$ in Equation (31).

w/o *f*. To show the enhancement of the multi-task joint learning loss function, we set α_1 and α_2 to 1 and it remained unchanged during the iterative training process.

We provide the ablation study results on the ESConv dataset in Table 6. From this table, we make observations as follows:

1. Mixed-Strategy Learning Module (Module u): Removal resulted in a significant decrease in various evaluation indicators, including ACC, D-1, B-2, R-L, Precision, and Recall by 3.4%, 16%, 64%, 14%, 8%, and 11%, which suggests that the Mixed-Strategy module can provide sufficient linguistic pattern and features references, and hence boost the performance of generating responses by specific strategy categories.
2. Improvement of the Emotional Reply Generation Stage (Module a): Removal led to varying degrees of decline in various evaluation indicators, dropping by 1.9%, 18.7%, 43%, 12.5%, 5%, and 6.7%. This demonstrates that the Response-Generate module has a notable impact on model performance, and it is necessary to consider the impact of multiple features during the response generation stage.
3. Multi-task Joint Learning Loss Function Improvement (Module *f*): Removal resulted in relatively small drops in evaluation metrics, with decreases of 1.4%, 10.7%, 18.9%, 4.7%, 4.3%, and 2.6%, respectively. It confirms that the dynamic multi-task joint learning loss function automatic average weight method has a positive effect on the model's overall performance.

Table 6. Ablation experiment results. The upward arrow in the figure indicates that the higher the evaluation standard, the better.

Model	ACC (%) ↑	D-1 ↑	B-2 ↑	R-L ↑	Precision ↑	Recall ↑
AdMISC	**34.41**	**3.83**	**8.16**	**17.91**	**28.88**	**31.25**
w/o *u*	33.22	3.20	2.92	15.27	26.46	27.81
w/o *a*	33.73	3.11	4.61	15.67	27.32	29.13
w/o *f*	33.90	3.42	6.61	17.06	27.61	30.45

5.6. Case Study

We illustrated a conversation in the test set to obtain an intuitive understanding of our model with some baselines, the results of which are presented in Table 7. Various problems appear in the compared models, such as inconsistency, repetition, contradiction, etc. The FADO model does not have obvious narrative errors, but it generates interrogative sentences, which are inappropriate for the emotional support task. Intuitively, AdMISC achieves the best performance in contrast.

Table 7. Comparison between AdMISC and baseline model responses (some contextual content has been ignored).

Background	
Seeker	My boyfriend and I recently broke up due to a long-distance relationship and the impact of COVID-19. . .
Context	
Seeker	Hi, I'm hoping you can help me get through this difficult time.
Supporter	Hi, can you please explain the difficult time you are having?
Seeker	My boyfriend broke up with me and I've been upset. We live in different parts of the country and. . .
	. . .. . .
Seeker	Have you ever gone through a breakup?
Response	
Transformer	I am so sorry that you are not able to get your friends and family.
MT Transformer	I have had a similar situation and I understand. It was the only way I was in a similar situation, I have found that way to get out of my life.
MoEL	I am so sorry that you are feeling better. I had a relationship with your boyfriend!
MIME	I know it is hard, but it is always good to do things out for yourself.
BlenderBot-Joint	I have not but i have been through a lot of breakups. . .
FADO (SOTA)	I understand that and it is sad. Have you tried to talk to him about it?
AdMISC	**I have been in a break up before and it was very tough. I understand that you have felt sad and you are feeling lonely.**
Ground Truth	**It sounds like you are really stressed over the breakup and you want to know how to move forward.**

5.7. Top-k Prediction and Label Prediction Results

We further demonstrated the effectiveness of our method by comparing Top-k prediction accuracy and visually comparing the distribution of strategy labels.

Top-k prediction. As shown in Figure 5, a comparison between AdMISC and the FADO model was conducted on Top-k prediction accuracy. AdMISC consistently outperforms the FADO model across all Top-k accuracy metrics, indicating more accurate label classification.

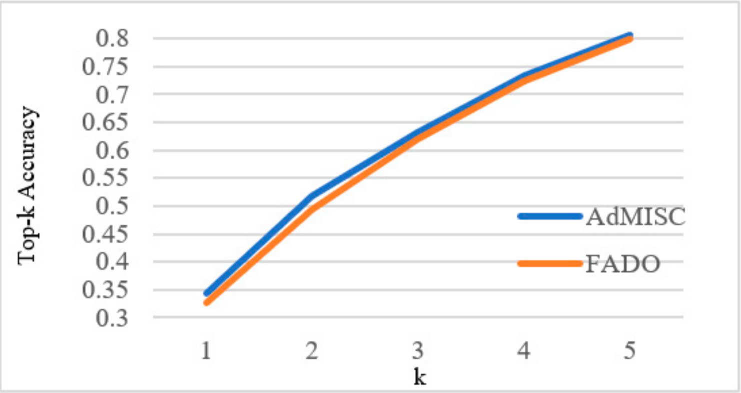

Figure 5. Top-K Accuracy between AdMISC and FADO.

Strategy label distribution. To deeply evaluate the performance of the AdMISC model in emotional label prediction, we compared the emotional label prediction results of FADO and AdMISC on the same problem and made statistics with the Ground-Truth emotional label, as shown in Figure 6.

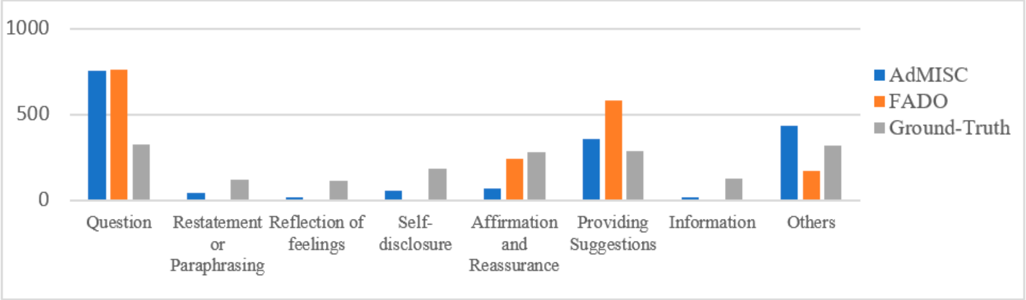

Figure 6. Label prediction distribution between FADO and AdMISC.

Analyzing the data revealed that FADO tended to classify more labels as "Question" in terms of emotion label prediction, resulting in significantly more labels than the Ground-Truth emotion labels, with only a handful of them correctly classified as "Self-disclosure". FADO's prediction results did not correctly classify any related issues in the four labels "Reflection of feelings", "Restatement or Paraphrasing", "Information", and "Others". In contrast, the label prediction distribution of the AdMISC model was more reasonable, with more labels correctly predicted, highlighting the positive role of the mixed strategy learning module in predicting strategy labels.

6. Conclusions

In this paper, we propose a multi-task joint learning method in emotionally supportive dialogue models. This method extracts dynamic emotional change information at the sentence level and combines commonsense-enhanced historical dialogue information and seeker's situation descriptions to guide strategy selection. Additionally, the method introduces multi-head attention mechanisms and cross-attention mechanism layers to enhance dialogue with the features extracted in feedforward networks, improving the quality of response generation. A series of experiments demonstrated the feasibility and effectiveness of this method. Furthermore, our approach may provide information for other downstream tasks in dialogue systems. For example, in open-domain dialogue systems or recommendation systems, strengthening the connection between contextually relevant information and target responses may allow the model to generate better quality responses. In future work, we will continue to explore other dialogue features that affect emotional support effects and implement our method using lighter weight networks.

7. Limitations

Although our method is a certain improvement over existing baseline models, we believe that there are still many issues that remain to be solved in the work of emotional support conversation models. First, the accuracy of our method in dialogue strategy prediction has improved, but it is still not high enough, and the model produced some errors at the prediction stage. One reason for these errors may be that the model needs more semantic information to help better establish the connection between context and supporting strategies. It also needs to build a larger corpus and clearer dialogue strategy annotations. Secondly, we used COMET to enhance the model with commonsense knowledge. However, more professional domain knowledge may be required for emotional support tasks, such as human health or mental health knowledge. In addition, we evaluated model performance using both automatic and human evaluation. However, the currently used automatic evaluation indicators are still not reasonable enough and cannot judge the emotional support ability of the model. Better evaluation indicators should be established for this purpose, and human evaluation methods are not professional enough. A completely different approach should probably be taken in strictly psychological research.

Author Contributions: Writing—original draft, X.J.; Supervision, J.H., Q.Z. and J.J. All authors have read and agreed to the published version of the manuscript.

Funding: Project of Sichuan Science and Technology Department (2021Z005).

Data Availability Statement: The data presented in this study are available in this article.

Conflicts of Interest: The author Qian Zhang was employed by the company Active Network (Chengdu), Ltd. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Green, Z.A.; Faizi, F.; Jalal, R.; Zadran, Z. Emotional support received moderates academic stress and mental well-being in a sample of Afghan university students amid COVID-19. *Int. J. Soc. Psychiatry* **2022**, *68*, 1748–1755. [CrossRef]
2. Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Health* **2017**, *4*, e19. [CrossRef]
3. Tu, Q.; Li, Y.; Cui, J.; Wang, B.; Wen, J.-R.; Yan, R. MISC: A MIxed Strategy-Aware Model Integrating COMET for Emotional Support Conversation. *arXiv* **2022**, arXiv:2203.13560.
4. Peng, W.; Hu, Y.; Xing, L.; Xie, Y.; Sun, Y.; Li, Y. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation. *arXiv* **2022**, arXiv:2204.12749.
5. Ding, Y.; Liu, J.; Zhang, X.; Yang, Z. Dynamic tracking of state anxiety via multi-modal data and machine learning. *Front. Psychiatry* **2022**, *13*, 757961. [CrossRef]
6. Zheng, C.; Liu, Y.; Chen, W.; Leng, Y.; Huang, M. CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation. *arXiv* **2021**, arXiv:2105.08316.
7. Wei, B.; Lu, S.; Mou, L.; Zhou, H.; Poupart, P.; Li, G.; Jin, Z. Why Do Neural Dialog Systems Generate Short and Meaningless Replies? A Comparison between Dialog and Translation. *arXiv* **2017**, arXiv:1712.02250.
8. Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA, 2018. [CrossRef]
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
10. Rui, Z. *Research on the Key Technologies of Emotion Aware Task Oriented Dialogue Strategy*; South China University of Technology: Guangzhou, China, 2021. [CrossRef]
11. Peng, W.; Qin, Z.; Hu, Y.; Xie, Y.; Li, Y. Fado: Feedback-aware double controlling network for emotional support conversation. *Knowl. Based Syst.* **2023**, *264*, 110340. [CrossRef]
12. Xu, X.; Meng, X.; Wang, Y. Poke: Prior knowledge enhanced emotional support conversation with latent variable. *arXiv* **2022**, arXiv:2210.12640.
13. Cheng, Y.; Liu, W.; Li, W.; Wang, J.; Zhao, R.; Liu, B.; Liang, X.; Zheng, Y. Improving Multi-turn Emotional Support Dialogue Generation with Lookahead Strategy Planning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; pp. 3014–3026.
14. Zeng, B.; Yang, H.; Xu, R.; Zhou, W.; Han, X. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Appl. Sci.* **2019**, *9*, 3389. [CrossRef]
15. Zhong, P.; Wang, D.; Li, P.; Zhang, C.; Wang, H.; Miao, C. CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts. *arXiv* **2020**, arXiv:2012.08377. [CrossRef]
16. Speer, R.; Chin, J.; Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
17. Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; Choi, Y. COMET: Commonsense transformers for automatic knowledge graph construction. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 4762–4779.
18. Deng, Y.; Zhang, W.; Yuan, Y.; Lam, W. Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations. *arXiv* **2023**, arXiv:2305.10172.

19. Weixiang, Z.; Yanyan, Z.; Shilong, W.; Qin, B. TransESC: Smoothing Emotional Support Conversation via Turn-Level State Transition. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 6725–6739.
20. Wang, C.; Ma, Z.; Du, B.; Jia, W.; Wang, H.; Bao, C. Survey of Research on End-to-End Emotional Dialogue Generation. *J. Front. Comput. Sci. Technol.* **2022**, *16*, 280–295.
21. Wang, H.; Wang, R.; Mi, F.; Deng, Y.; Wang, Z.; Liang, B.; Xu, R.; Wong, K.-F. Chain-of-thought prompting for responding to in-depth dialogue questions with LLM. *arXiv* **2023**, arXiv:2305.11792.
22. Yangzhou; Chen, Z.; Cai, T.; Wang, Y.; Liao, X. A Review of Emotional Dialogue Response Based on Deep Learning. *J. Comput. Sci.* **2023**, *46*, 2489–2519.
23. Hill, C.E. *Helping Skills: Facilitating, Exploration, Insight, and Action*; American Psychological Association: Washington, DC, USA, 2009.
24. Liu, B.; Sundar, S.S. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychol. Behav. Soc. Netw.* **2018**, *21*, 625–636. [CrossRef]
25. Li, Q.; Li, P.; Ren, Z.; Chen, Z. Knowledge bridging for empathetic dialogue generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 22 February–1 March 2022; AAAI Press: Washington, DC, USA, 2022; Volume 36, pp. 10993–11001.
26. Kim, T.; Vossen, P. Emoberta: Speaker-aware emotion recognition in conversation with Roberta. *arXiv* **2021**, arXiv:2108.12009.
27. Peng, W.; Hu, Y.; Yu, J.; Xing, L.; Xie, L. APER: Adaptive evidence-driven Reasoning Network for machine reading comprehension with unanswerable questions. *Knowl. Based Syst.* **2021**, *229*, 107364. [CrossRef]
28. Liu, S.; Johns, E.; Davison, A.J. End-to-End Multi-Task Learning with Attention. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019. [CrossRef]
29. Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; Huang, M. Towards emotional support dialog systems. *arXiv* **2021**, arXiv:2106.01144.
30. Xu, J.; Ju, D.; Li, M.; Boureau, Y.-L.; Weston, J.; Dinan, E. Recipes for Safety in Open-domain Chatbots. *arXiv* **2020**, arXiv:2010.07079.
31. Rashkin, H.; Smith, E.M.; Li, M.; Boureau, Y.-L. I know the feeling: Learning to converse with empathy. *arXiv* **2018**, arXiv:1811.00207.
32. Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; Fung, P. 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), Hong Kong, China, 3–7 November 2019. [CrossRef]
33. Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A.; Mihalcea, R.; Poria, S. *MIME: MIMicking Emotions for Empathetic Response Generation*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.