Multi-stream Information Fusion Framework for Emotional Support Conversation

Yinan Bao[1,2], Dou Hu[1,2], Lingwei Wei[1,2], Shuchong Wei[1,2], Wei Zhou[1,*], Songlin Hu[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{baoyinan, hudou, weilingwei, weishuchong, zhouwei, husonglin}@iie.ac.cn

Abstract

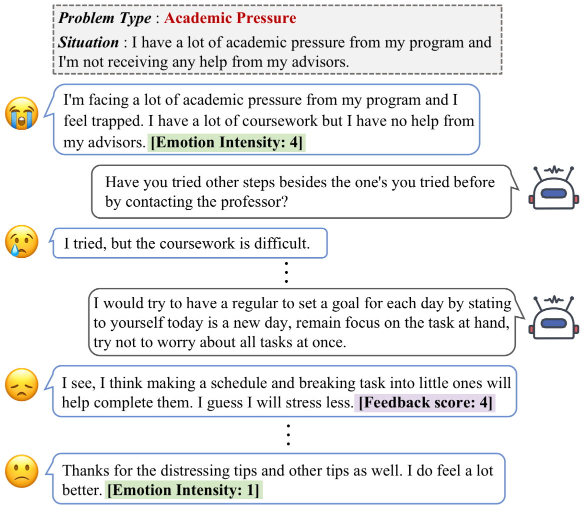
Emotional support conversation (ESC) task aims to relieve the emotional distress of users who have high-intensity of negative emotions. However, due to the ignorance of emotion intensity modelling which is essential for ESC, previous methods fail to capture the transition of emotion intensity effectively. To this end, we propose a **M**ulti-stream information **F**usion **F**ramework (**MFF-ESC**) to thoroughly fuse three streams (text semantics stream, emotion intensity stream, and feedback stream) for the modelling of emotion intensity, based on a designed multi-stream fusion unit. As the difficulty of modelling subtle transitions of emotion intensity and the strong emotion intensity-feedback correlations, we use the Kullback–Leibler divergence between feedback distribution and emotion intensity distribution to further guide the learning of emotion intensities. Experimental results on automatic and human evaluations indicate the effectiveness of our method.

Keywords: emotional support conversation, emotion intensity, response generation

1. Introduction

Emotional support conversation (ESC) aims to provide effective support for users who have high-intensity of negative emotions (Burleson et al., 2006; Heaney and Israel, 2008; Slovák et al., 2015) and reduce their emotional distress as the dialogue goes (Liu et al., 2021a). Due to its potential applications to provide users with in-time emotional support in social interactions, mental health support, and so on (Liu et al., 2021a; Peng et al., 2022), researchers have shown increasing attention to ESC.

Unlike traditional emotional tasks, ESC system should perceive the transition of user's emotion intensity and generate supportive responses accordingly to decrease the intensity. However, existing methods (Liu et al., 2021a; Tu et al., 2022; Peng et al., 2022; Zhao et al., 2023b) overlook the modelling of emotion intensity due to the difficulty to model its dynamic transition.



Problem Type : Academic Pressure
Situation : I have a lot of academic pressure from my program and I'm not receiving any help from my advisors.

I'm facing a lot of academic pressure from my program and I feel trapped. I have a lot of coursework but I have no help from my advisors. **[Emotion Intensity: 4]**

Have you tried other steps besides the one's you tried before by contacting the professor?

I tried, but the coursework is difficult.

⋮

I would try to have a regular to set a goal for each day by stating to yourself today is a new day, remain focus on the task at hand, try not to worry about all tasks at once.

I see, I think making a schedule and breaking task into little ones will help complete them. I guess I will stress less. **[Feedback score: 4]**

⋮

Thanks for the distressing tips and other tips as well. I do feel a lot better. **[Emotion Intensity: 1]**

Figure 1: An example of an emotional support conversation between user and the system.

To this end, we propose to fuse three kinds of streams for the effective modelling of the dynamic transition of emotion intensity. As shown in Figure 1, there are three main streams throughout the conversation, including the text semantics stream, feedback stream, and emotion intensity stream. Any two of the three kinds of streams have correlations. **First**, the semantics conveyed by the text content reflects user's emotion intensity and implies user's feedback about the system. **Second**, the feedback scores affects users' following semantic expression and indirectly reflects the subtle fluctuation of emotion intensity. **Third**, the transition of emotion intensity can be mapped to the change of feedback score and influence users' expression as well. To fuse the three streams sufficiently, we subtly design a multi-stream fusion unit to model the interactions among the three streams.

However, it is an intractable challenge for modelling the transition of emotion intensity effectively as the subtle change of emotion intensity is difficult to capture. Apparently, feedback score indicates how much user's emotional distress is relieved and user's emotion intensity should decrease with the increase of feedback score. To alleviate this problem, we fully utilize the strong correlation between feedback and emotion intensity. Specifically, we use a transformation to project the negative corre-

[*] Corresponding author.

LREC-COLING 2024, pages 11981–11992
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

lation to a positive correlation. Then, we pull in the feedback distribution and emotion intensity distribution under the supervision of Kullback–Leibler (KL) divergence, guiding the representation learning of emotion intensities in the latent space. In addition, we utilize the true labels of feedback score and emotion intensity to supervise the representations generated by the proposed multi-stream fusion unit.

Apart from perceiving the transition of emotion intensity, it's important as well to distil crucial cause cues about emotional distress from the noisy conversation history. The prior knowledge about problem type of users in the whole corpus has potential to comprehend users' explicit problems and further capture implicit cause cues related to the specific problem type from the dialogue history.

In this paper, we propose **M**ulti-stream Information **F**usion **F**ramework for **ESC** (MFF-ESC), fusing three sequential streams thoroughly based on a novel **M**ulti-stream **F**usion **U**nit (MFU) for better perception of the fluctuation of emotion intensity and realizing problem type semantics' potential to explore crucial cause cues from noisy context. For the subtly designed MFU, we modify the one hidden state of LSTM (Hochreiter and Schmidhuber, 1997) unit to two, representing the state of feedback stream and emotion intensity stream respectively. For the injection of problem type information, we implement cross-attention to model problem type-situation and problem type-context interaction, obtaining cause-related semantics. Experimental results indicate that MFF-ESC improves the accuracy and diversity of generated responses and outperforms state-of-the-art methods. Our contributions are as follows:

- We make the first attempt to model the transition of emotion intensity for ESC, based on a novel designed multi-stream fusion unit for the thorough fusion of three streams (text semantics stream, feedback stream, emotion intensity stream).
- Due to the difficulty of emotion intensity modelling and the strong emotion intensity-feedback correlations, we adopt KL divergence to minimize the distance between feedback distribution and emotion intensity distribution, further guiding the learning of emotion intensities.
- Experiments on the benchmark dataset demonstrate the superiority of MFF-ESC, compared with state-of-the-art methods.

2. Related Work

2.1. Empathetic Conversation System

Emotional conversation systems have attracted increasing interest in recent years. Empathetic response generation (Zheng et al., 2021; Majumder et al., 2020a; Zhong et al., 2020; Chen et al., 2022) is a typical task that aims to understand users' feelings and then reply accordingly in an empathetic way. For an empathetic dialogue, users may have positive emotions like happiness or negative emotions like sadness and the system needs to recognize the emotion and caters to it. However, ESC only focuses on users with high-intensity of negative emotions and needs to understand uses' problem, perceive the transition of emotion intensity, and generate responses to decrease the intensity.

2.2. Emotional Support Conversation System

ESC is a newly proposed text generation task that aims to reduce users' emotional distress by providing supportive responses (Liu et al., 2021a). Tu et al. (2022) utilized commonsense knowledge to enhance the interaction modelling between user's situation and dialogue history. Peng et al. (2022) used the commonsense knowledge to obtain the psychological intent of users and utilized problem type to supervise the representation learning of situation. Cheng et al. (2022) proposed a lookahead strategy planning with foresight of the user feedback to select strategies that are beneficial from a long-term perspective. Cheng et al. (2023) introduced users' persona information into ESC, providing personalized emotional support. Zhou et al. (2023) used COMET (Bosselut et al., 2019) and VAD (Mohammad, 2018) to track the transition of users' emotional states. Zhao et al. (2023b) modelled turn-level emotion transition on the basis of a pre-trained model for emotion prediction. However, the external emotional information introduced by Zhou et al. (2023) and Zhao et al. (2023b) contains emotions with different polarities, which is different from the negative emotion intensity focused by ESC.

Existing methods ignore the modelling of emotion intensity which is essential for ESC. In this paper, we aim to fully utilize emotional information within the dataset and model the transition of emotion intensity with the assistance of text semantics stream and feedback stream.

3. Preliminaries

3.1. ESConv Dataset

Given a conversation between a user and the system and user's side information (emotion intensity, feedback, situation, problem type), ESC aims to generate a response for the reduction of user's negative emotion intensity. We use the benchmark dataset ESConv (Liu et al., 2021a) for experiments.

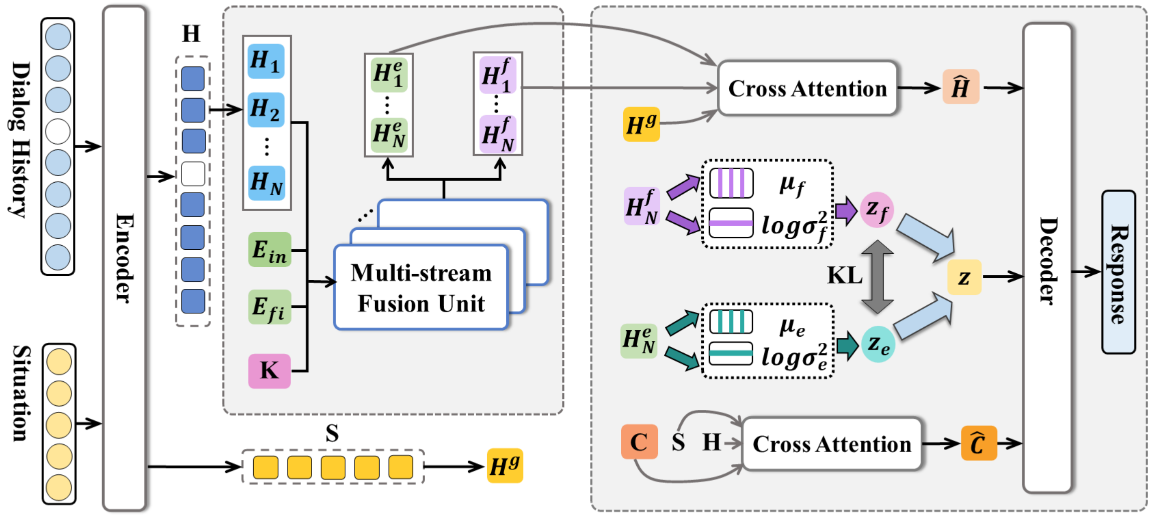


Figure 2: An overview of MFF-ESC. **Dialogue History** means the context before the current step. **Situation** means the brief description of user's situation before the dialogue. **C** means the embedding of user's problem type. $\mathbf{E}_{in}$ and $\mathbf{E}_{fi}$ means user's emotion intensity at the beginning and end of the dialogue. $\mathbf{K}$ means the sequence of user's feedback scores for the dialogue history. $\mathbf{z}$ presents $\mathbf{z}_e$ during the training process and $\mathbf{z}_f$ during the testing process, because user's true emotion intensity at the end of a conversation is unknown during the testing process.

Apart from the text content, each conversation contains user's emotion intensity at the beginning and end of the conversation and user's feedback score after every two utterances he/she received from the system[1]. Before a conversation, some prior information related to user's emotional distress is provided, making the system aware of the difficulty that user faces. The prior information includes the problem type of user's emotional distress and a brief introduction of user's situation.

3.2. Problem Formulation

Given a dialogue history $D = (u_1, u_2, ..., u_N)$ that consists of N utterances, the user's emotion intensity $\mathbf{e}_{in}$ and $\mathbf{e}_{fi}$ at the beginning and end of the conversation respectively, the user's feedback score $K = (k_1, k_2, ..., k_m)$[2], the corresponding problem type C, and the global situation $\mathbf{s} = (s_1, s_2, ..., s_{|S|})$ with $|S|$ words, the target of ESC is to generate a supportive response $\mathbf{r}$ to decrease user's negative emotion intensity. In conclusion, the target is to estimate the probability distribution $\mathbf{p}(\mathbf{r}|D, \mathbf{e}_{in}, \mathbf{e}_{fi}, K, C, \mathbf{s})$.

4. Methodology

4.1. Context Encoder

Following existing methods (Liu et al., 2021a; Tu et al., 2022; Peng et al., 2022), we encode the conversation history and situation of user based on the encoder of Blenderbot (Roller et al., 2021).

$$\begin{aligned} \mathbf{H} &= \mathbf{Enc}([CLS], u_1, [SEP], ..., u_N, [SEP])\,, \\ \mathbf{H}^l &= \texttt{max-pooling}(\mathbf{H})\,, \end{aligned} \tag{1}$$

where $\mathbf{H} \in \mathbb{R}^{T_l \times d}$ presents the representation of the input sequence with T_l tokens, $\mathbf{H}^l = (\mathbf{H}_1, ..., \mathbf{H}_N)$, and $\mathbf{H}_i \in \mathbb{R}^d$ is the representation of utterance u_i.

For user's situation, the token-level vector $\mathbf{S} \in \mathbb{R}^{T_g \times d}$ and sentence-level vector $\mathbf{H}^g \in \mathbb{R}^d$ are obtained in a similar way, where T_g means the number of tokens of situation.

4.2. Multi-stream Fusion Unit

To model the transition of emotion intensity, we subtly design a **M**ulti-stream **F**usion **U**nit (MFU) based on LSTM (Hochreiter and Schmidhuber, 1997) to fuse three streams thoroughly, including text semantics stream, emotion intensity stream, and feedback stream. The excellent abilities of long-term and short-term memory of LSTM motivate us to choose it as the basis of MFU.

As shown in Figure 3, MFU contains forget gate $(\mathbf{f}_t)$, input gate $(\mathbf{i}_t)$, and two output gates $(\mathbf{o}^e_t, \mathbf{o}^f_t)$. The difference between MFU and LSTM is MFU has two output gates for calculating the hidden

[1] The scores of feedback and emotion intensity are integers ranging from 1 to 5.

[2] $m = N_s/2$ refers to the number of feedback scores in a conversation, where N_s means the number of system's utterances in a conversation.

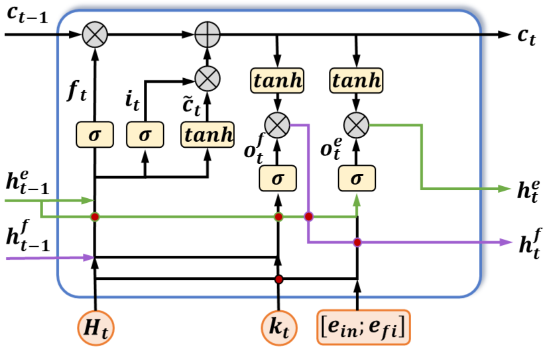


Figure 3: The structure of Multi-stream Fusion Unit. Red cross points means the vectors are not concatenated.

states of emotion intensity and feedback respectively. The calculation of gates in MFU is as follows:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}^e; \mathbf{h}_{t-1}^f; \mathbf{H}_t] + \mathbf{b}_f)\,, \\
\mathbf{i}_t &= \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}^e; \mathbf{h}_{t-1}^f; \mathbf{H}_t] + \mathbf{b}_i)\,, \\
\tilde{\mathbf{c}}_t &= \mathtt{tanh}(\mathbf{W}_c[\mathbf{h}_{t-1}^e; \mathbf{h}_{t-1}^f; \mathbf{H}_t] + \mathbf{b}_c)\,, \\
\mathbf{o}_t^e &= \sigma(\mathbf{W}_o^e[\mathbf{h}_{t-1}^e; \mathbf{e}_{in}; \mathbf{e}_{fi}; \mathbf{H}_t] + \mathbf{b}_o^e)\,, \\
\mathbf{o}_t^f &= \sigma(\mathbf{W}_o^f[\mathbf{h}_{t-1}^f; \mathbf{k}_t; \mathbf{H}_t] + \mathbf{b}_o^f)\,,
\end{aligned} \tag{2}$$

where the forget gate $\mathbf{f}_t$ and input gate $\mathbf{i}_t$ are computed based on the input of text semantics stream $\mathbf{H}_t$ and the hidden states of emotion intensity stream and feedback stream. $\mathbf{e}_{in}$ and $\mathbf{e}_{fi}$ mean user's emotion intensity at the beginning and end of the conversation, respectively. $\mathbf{k}_t$ means the feedback score at the step t. $\tilde{\mathbf{c}}_t$ is the updated cell state. $\mathbf{o}_t^e$ and $\mathbf{o}_t^f$ mean the output gate of emotion intensity stream and feedback stream, respectively [3].

The calculations of cell state and hidden states in MFU are as follows:

$$\begin{aligned}
\mathbf{c}_t &= \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{c}}_t\,, \\
\mathbf{h}_t^e &= \mathbf{o}_t^e * \mathtt{tanh}(\mathbf{c}_t)\,, \\
\mathbf{h}_t^f &= \mathbf{o}_t^f * \mathtt{tanh}(\mathbf{c}_t)\,.
\end{aligned} \tag{3}$$

Given a conversation with semantic stream $\mathbf{H}^l = (\mathbf{H}_1, ..., \mathbf{H}_N)$, emotion intensity stream $\mathbf{E} = (\mathbf{e}_{in}, \mathbf{e}_{fi})$, and feedback stream $\mathbf{K} = (\mathbf{k}_1, ..., \mathbf{k}_m)$, we obtain two hidden state sequences: $\mathbf{H}^e = (\mathbf{h}_1^e, ..., \mathbf{h}_N^e)$ and $\mathbf{H}^f = (\mathbf{h}_1^f, ..., \mathbf{h}_N^f)$ by using the MFU recurrently.

As shown in Equation (2), we only incorporate emotion intensity stream and feedback stream in the output gate rather than all the gates. The reason is that emotion intensity and feedback are discrete scores with limited semantics compared with text content with rich semantics, the semantic spaces of these two streams and $\mathbf{H}^l$ are much more different. To trade off the dialogue history comprehension and the injection of emotion intensity and feedback, we only use these two streams to directly influence the calculation of output gates which will reflect on $\mathbf{h}_t^e$ and $\mathbf{h}_t^f$ with short-term memory. The cell state $\mathbf{c}_t$ with long-term memory won't be directly influenced but contains certain information about emotion intensity and feedback streams as well [4].

4.3. Response Generator

In the response generator, we inject emotion intensity, user's situation information, and problem type of user to facilitate the response generation based on the decoder of Blenderbot (Roller et al., 2021).

Injection of Emotion Intensity. As the annotation only contains user's emotion intensity at the beginning and end of the conversation, to further supervise the learning of emotion intensity, we use KL divergence to minimize the distance between the distributions of emotion intensity and feedback. Then, we sample a latent variable $\mathbf{z}$ from one of the two distributions for response generation. Detailed operations are as follow.

We assume $\mathbf{z}$ follows isotropic Gaussian distribution. Taking the vector $\mathbf{H}_N^e$ of the last utterance as input, for the approximate posterior distribution $q_\theta(z|D, \mathbf{e}_{in}, \mathbf{e}_{fi}, K) \sim \mathcal{N}(\mu_e, \sigma_e^2\mathbf{I})$, we obtain μ_e and $\log\sigma_e^2$ as follows:

$$\mu_e, \log\sigma_e^2 = \mathbf{FNN}^e(\mathbf{H}_N^e)\,, \tag{4}$$

where $\mathbf{FNN}^e$ means a three-layer feed-forward neural network. We use the reparameterization trick (Kingma and Welling, 2014) to sample a latent variable $\mathbf{z}_e \in \mathbb{R}^{d_l}$ from $q_\theta(z|D, \mathbf{e}_{in}, \mathbf{e}_{fi}, K)$.

Similarly, we take the vector $\mathbf{H}_N^f$ of the last utterance as input, using a $\mathbf{FNN}$ to get μ_f and $\log\sigma_f^2$ of the approximate posterior distribution $p_\phi(z|D, \mathbf{e}_{in}, K) \sim \mathcal{N}(\mu_f, \sigma_f^2\mathbf{I})$ and sample a latent variable $\mathbf{z}_f \in \mathbb{R}^{d_l}$.

Injection of Situation. To integrate user's situation with informative cause cues of distress and highlight the semantics implied in the last utterance $\mathbf{H}_N^e$, we use cross attention mechanism to model the interaction between $\mathbf{H}^g$ and $\mathbf{H}^e$ as follows:

$$\begin{aligned}
\hat{\mathbf{H}}_e^g &= \mathtt{cross\text{-}att}(\mathbf{H}^g, \mathbf{H}^e)\,, \\
\hat{\mathbf{H}}_N^e &= \mathtt{cross\text{-}att}(\mathbf{H}_N^e, \mathbf{H}^e)\,.
\end{aligned} \tag{5}$$

In the same way, $\hat{\mathbf{H}}_f^g$ and $\hat{\mathbf{H}}_N^f$ are obtained by the interaction modelling between $\mathbf{H}^g$ and $\mathbf{H}^f$. Then,

[3] In the testing process, we replace $\mathbf{e}_{fi}$ with the learned embedding of the lowest emotion intensity to calculate $\mathbf{o}_t^e$.

[4] We have compared variants of MFU in the experiments.

we obtain the vector $\hat{\mathbf{H}}$ integrated with situation information as follows:

$$\begin{aligned} \hat{\mathbf{H}}^{g} &= \mathbf{FNN}([\hat{\mathbf{H}}_{e}^{g}; \hat{\mathbf{H}}_{f}^{g}])\,, \\ \hat{\mathbf{H}}_{N} &= \mathbf{FNN}([\hat{\mathbf{H}}_{N}^{e}; \hat{\mathbf{H}}_{N}^{f}])\,, \\ \hat{\mathbf{H}} &= \mathtt{tanh}(\hat{\mathbf{H}}^{g} + \hat{\mathbf{H}}_{N})\,, \end{aligned} \tag{6}$$

where $\mathbf{FNN}$ means a two-layer feed-forward neural network with $\mathbf{ReLU}$ as the activation function.

Injection of Problem Type. To incorporate problem type information with high-level causal semantics, we explore the explicit and implicit cues implied in situation and dialogue history respectively to update the representation of problem type. We adopt the following methods to obtain explicit semantic-enhanced cause embedding $\mathbf{C}_{ex}$:

$$\begin{aligned} \alpha_{ex} &= \mathbf{W}_{2}[\mathbf{C}\,;\, (\mathbf{S}\mathbf{W}_{1} + \mathbf{b}_{1})]^{\top} + \mathbf{b}_{2}\,, \\ \alpha_{ex} &= \mathtt{softmax}(\alpha_{ex})\,, \\ \mathbf{C}_{ex} &= \mathbf{FNN}([\alpha_{ex}\mathbf{S}\,;\, \mathbf{C}])\,, \end{aligned} \tag{7}$$

where $\mathbf{C}$ is the embedding of problem type; $\mathbf{W}_1 \in \mathbb{R}^{d\times d}, \mathbf{W}_2 \in \mathbb{R}^{d+d_c}$ are model parameters; $\alpha_{ex} \in \mathbb{R}^{T_g}$ measures the importance of each token in the situation, with the guidance of problem type.

To grasp the implicit cues from the noisy dialogue history, we extract keywords from the context by TextRank algorithm (Mihalcea and Tarau, 2004). Then, we use the keyword vectors to generate the implicit semantic-enhanced vector of problem type:

$$\begin{aligned} \alpha_{im} &= \mathbf{W}_{4}[\mathbf{C}\,;\, (\mathbf{H}\mathbf{W}_{3} + \mathbf{b}_{3})]^{\top} + \mathbf{b}_{4}\,, \\ \alpha_{im} &= \mathtt{softmax}(\mathbf{M} \odot \alpha_{im})\,, \\ \mathbf{C}_{im} &= \mathbf{FNN}([\alpha_{im}\mathbf{H}\,;\, \mathbf{C}])\,, \end{aligned} \tag{8}$$

where $\odot$ is the element-wise product operation; $\mathbf{W}_3 \in \mathbb{R}^{d\times d}, \mathbf{W}_4 \in \mathbb{R}^{d+d_c}$ are model parameters; $\mathbf{M} \in \mathbb{R}^{T_l}$ masks the tokens that aren't keywords in the dialogue history.

Then, we fuse $\mathbf{C}_{ex}$ and $\mathbf{C}_{im}$ to obtain the final problem type embedding:

$$\hat{\mathbf{C}} = \mathtt{tanh}(\mathbf{C}_{ex} + \mathbf{C}_{im})\,. \tag{9}$$

Finally, we generate the response based on the decoder of BlenderBot, following existing methods (Liu et al., 2021a; Tu et al., 2022; Peng et al., 2022). Detailed operations are as follows:

$$\begin{aligned} &\mathbf{O} = \mathbf{FNN}([\hat{\mathbf{H}}\,;\, \mathbf{z}\,;\, \hat{\mathbf{C}}])\,, \\ &\alpha = \mathtt{softmax}(\mathbf{W}_{5}[\mathbf{O} \odot \mathbf{H}]^{\top} + \mathbf{b}_{5})\,, \\ &\mathbf{p}(\mathbf{r}|D, \mathbf{e}_{in}, \mathbf{e}_{fi}, K, C, \mathbf{s}) = \mathbf{Dec}(\alpha\mathbf{O} + \mathbf{H})\,, \end{aligned} \tag{10}$$

where $\mathbf{z}$ presents $\mathbf{z}_e$ during training process and $\mathbf{z}_f$ during testing process; $\mathbf{O} \in \mathbb{R}^d$ is the representation after feature transformation; $\mathbf{W}_5 \in \mathbb{R}^d$ is model parameter; $\alpha \in \mathbb{R}^{T_l}$ is the attention score to aggregate $\mathbf{O}$; $\mathbf{Dec}$ is the decoder of BlenderBot.

4.4. Loss Function

The reconstruction loss of a response with N_r tokens is:

$$\mathcal{L}_r = -\sum_{t=1}^{N_r} \mathtt{log}\,\mathbf{p}(r_t|r_{j<t}, D, \mathbf{e}_{in}, \mathbf{e}_{fi}, K, C, \mathbf{s})\,. \tag{11}$$

We use the true label of initial emotion intensity, final emotion intensity, and feedback to supervise the learning of $\mathbf{H}_1^e$, $\mathbf{H}_N^e$, and $\mathbf{H}^f$ in MFF-ESC, obtaining $\mathcal{L}_{in}, \mathcal{L}_{fi}$, and $\mathcal{L}_f$.

$$\mathcal{L}_{MFF} = \mathcal{L}_{in} + \mathcal{L}_{fi} + \mathcal{L}_f\,. \tag{12}$$

Besides, the emotion intensity vector is further supervised by feedback vector with KL divergence. Notably, the original emotion intensity-feedback correlation is negative, as the emotion intensity should decline with the increase of feedback score. We argue that the $\mathbf{FNN}^e$ in Equation (4) will project the negative correlation to a positive correlation and adopt KL divergence to minimize the distance between the two distributions:

$$\mathcal{L}_{KL} = \mathbf{KL}(q_\theta(z|D, \mathbf{e}_{in}, \mathbf{e}_{fi}, K)\,||\,p_\phi(z|D, \mathbf{e}_{in}, K))\,. \tag{13}$$

Following existing methods (Tu et al., 2022; Peng et al., 2022), we use the true label of problem type and strategy to supervise $\hat{\mathbf{C}}$ and $\hat{\mathbf{H}}$, obtaining $\mathcal{L}_c$ and $\mathcal{L}_{st}$.

$$\mathcal{L}_{other} = \mathcal{L}_c + \mathcal{L}_{st}\,. \tag{14}$$

The final loss function $\mathcal{L}$ is:

$$\mathcal{L} = \lambda_1\mathcal{L}_r + \lambda_2\mathcal{L}_{MFF} + \lambda_3\mathcal{L}_{KL} + \lambda_4\mathcal{L}_{other}\,. \tag{15}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyper-parameters.

5. Experimental Settings

5.1. Dataset

Following previous methods (Liu et al., 2021a; Tu et al., 2022; Peng et al., 2022; Zhao et al., 2023b), we conduct experiments on the ESConv dataset collected by Liu et al. (2021a), which is based on English and contains 1,300 dialogues and 38,365 utterances. Each dialogue has 29.8 utterances on average. Following the original division, we split the dataset into the train, validation, and test sets with a ratio of 8:1:1.

6. Implementation Details

Following existing works (Liu et al., 2021a; Tu et al., 2022; Peng et al., 2022), we implement our method

Method	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	D-1↑	D-2↑	R-L↑
Transformer (Vaswani et al., 2017)	81.55	17.25	5.66	2.32	1.31	1.25	7.29	14.68
MoEL (Lin et al., 2019)	62.93	16.02	5.02	1.90	1.14	2.71	14.92	14.21
MIME (Majumder et al., 2020b)	43.27	16.15	4.82	1.79	1.03	2.56	12.33	14.83
DialoGPT-Joint (Liu et al., 2021a)	19.41	17.06	6.22	2.87	1.57	2.82	17.30	15.03
BlenderBot-Joint (Liu et al., 2021a)	16.11	17.27	6.33	3.17	1.81	3.60	21.88	15.20
MISC (Tu et al., 2022)	16.32	17.73	6.75	3.23	1.83	4.19	17.76	15.43
GLHG† (Peng et al., 2022)	15.67	19.66	7.57	3.74	2.13	3.50	21.61	16.37
MultiESC‡ (Cheng et al., 2022)	-	19.02	8.37	4.50	2.69	-	-	17.04
SUPPORTER (Zhou et al., 2023)	**15.39**	18.05	6.80	3.20	1.71	4.94	27.81	16.85
PAL (Cheng et al., 2023)	16.78	18.77	6.91	3.03	1.51	4.10	22.73	15.29
TransESC (Zhao et al., 2023b)	15.85	17.08	7.18	3.78	2.28	4.67	20.91	17.30
LLaMA-7B (0 shot)	-	4.79	2.00	0.99	0.52	2.82	17.21	8.51
ChatGPT (1 shot)§	-	13.91	4.53	1.96	1.02	5.92	31.38	13.19
ChatGLM-6B w/ P-Tuning	-	17.75	7.22	3.78	2.12	**7.46**	**35.00**	16.15
MFF-ESC (ours)	16.43	**20.64**	**8.87**	**4.81**	**2.98**	5.34	22.18	**18.83**

Table 1: Results of automatic metrics. † means the results are cited from the corresponding paper. ‡ means that we replace the backbone of MultiESC with Blenderbot-small for a fair comparison. § means the results are cited from Zhao et al. (2023a). Other baselines' results are reproduced based on the open-source codes. The results of baselines and our method are all based on the average score of 3 random runs. The bold results are the best and the underlined results are the second-best.

based on the small version of BlenderBot (Roller et al., 2021). We train MFF-ESC on a Tesla-V100 GPU with an initial learning rate of 2e-5 and a linear warmup of 100 steps. The batch size of training is 20. λ_1, λ_3 are 1 and λ_2, λ_4 are 0.5. d_c , d_l, and d are 50, 100, 512. The size of hidden states in MFU is 200. We train MFF-ESC for 5 epochs and obtain the decoded responses by Top-p and Top-k sampling with p = 0.6, k =30, temperature = 0.7, and repetition penalty = 1.03.

6.1. Baselines

We divide the competitive baselines into two categories, including baselines for empathetic responding and baselines for ESC. 1) Empathetic responding baselines: **Transformer** (Vaswani et al., 2017), **MoEL** (Lin et al., 2019), **MIME** (Majumder et al., 2020b). 2) ESC baselines: **DialoGPT-Joint** (Liu et al., 2021a), **BlenderBot-Joint** (Liu et al., 2021a), **MISC** (Tu et al., 2022), **GLHG** (Peng et al., 2022), **MultiESC** (Cheng et al., 2022), **SUPPORTER** (Zhou et al., 2023), **PAL** (Cheng et al., 2023), **TransESC** (Zhao et al., 2023b), **LLaMA-7B (0 shot)** (Touvron et al., 2023), **ChatGPT (1 shot)** (Zhao et al., 2023a), **ChatGLM-6B** (Du et al., 2022) with P-Tuning (Liu et al., 2021b).

6.2. Evaluation Metrics

Automatic Evaluation. Following existing methods (Peng et al., 2022; Zhao et al., 2023b), we use **PPL** (perplexity) to evaluate the general quality of the model. **B-1** (BLEU-1), **B-2** (BLEU-2), **B-3** (BLEU-3), **B-4** (BLEU-4) (Papineni et al., 2002), and **R-L** (ROUGE-L) (Lin, 2004) are adopted to measure the lexical and semantic aspects of the responses. **D-1** (Distinct-1) and **D-2** (Distinct-2) (Li et al., 2016) are utilized to evaluate the diversity of responses.

Comparisons	Aspects	Win	Lose	Tie
MFF-ESC vs. BlenderBot-Joint	Flu.	**45.6**	3.1	51.3
	Ide.	**57.5**	10.4	32.1
	Com.	**58.1**	11.5	30.4
	Sug.	**41.7**	8.3	50.0
	Ove.	**55.4**	12.5	32.1
MFF-ESC vs. TransESC	Flu.	**47.6**	13.8	38.6
	Ide.	**60.1**	11.7	28.2
	Com.	**52.3**	9.8	37.9
	Sug.	**45.8**	12.1	42.1
	Ove.	**62.9**	15.6	21.5

Table 2: Human evaluation results (%), which have significant improvement with p-value < 0.05.

Human Evaluation. We sample 100 instances for human evaluation. Given the generated responses of our method and a compared baseline model, we recruit three graduate students with linguistic or psychological background to select the better one. Following Liu et al. (2021a), we ask the annotators to choose the better model from five aspects: 1) **Fluency (Flu.)**: which model generates more fluent responses? 2) **Identification (Ide.)**: which model identify your problems better? 3) **Comforting (Com.)**: which model is more skillful in comforting you? 4) **Suggestion (Sug.)**: which model provides more helpful suggestions for you? 5) **Overall (Ove.)**: which model's emotional support do you prefer overall?

Method	B-1↑	B-2↑	B-4↑	R-L↑
MFF-ESC	**20.64**	**8.87**	**2.98**	**18.83**
- w/o Initial Emotion	19.03	7.65	2.39	17.89
- w/o Final Emotion	19.05	7.75	2.24	17.69
- w/o Feedback	19.04	7.78	2.28	17.60
- w/o Problem Type	19.61	7.99	2.50	18.09
- w/o $\mathcal{L}_{in}$	19.52	8.13	2.73	17.85
- w/o $\mathcal{L}_{fi}$	19.32	7.98	2.75	18.34
- w/o $\mathcal{L}_{f}$	18.75	7.84	2.59	17.75
- w/o $\mathcal{L}_{KL}$	18.43	7.18	2.15	17.90
- w/o $\mathcal{L}_{c}$	19.51	7.86	2.48	18.34
- w/o $\mathcal{L}_{st}$	19.85	8.16	2.43	17.31

Table 3: Evaluation results of ablation study.

7. Results and Analysis

7.1. Overall Results

Automatic Evaluation. As shown in Table 1, **MFF-ESC** outperforms the other baselines with the same backbone on most of the metrics. The promotions of **B-n**, **D-n**, and **R-L** show the effectiveness of our method.

Moreover, **MFF-ESC** which hasn't used any external knowledge even surpasses the state-of-the-art methods (GLHG, MultiESC, SUPPORTER, TransESC) which have utilized external knowledge to enhance context comprehension, proving its superiority. The results indicate that with the thorough fusion of text semantics stream, feedback stream, and emotion intensity stream based on MFU, **MFF-ESC** can perceive the transition of user's emotion intensity exquisitely and further contribute to the response generation.

Human Evaluation. As shown in Table 2, the results are consistent with those of automatic metrics. Compared with **BlenderBot-Joint**, **MFF-ESC** outperforms the most on **Ide.**, **Com.**, and **Ove.**, indicating that our method can understand user's problem better with the perception of emotion intensity, and finally leading to the generated responses with better empathy. For **TransESC** with external knowledge, **MFF-ESC** still gains more preferences on all of the five aspects. It proves that modeling emotion intensity is conducive to the generation of supportive responses.

7.2. Ablation Study

To verify the effects of different modules in our method, we conduct a series of ablation studies. The ablation study is divided into two parts. As shown in Table 3, the upper part means eliminating the corresponding information completely in the modelling process. The lower part means only removing the relative loss of the corresponding information.

Method	B-1↑	B-2↑	B-3↑	B-4↑	R-L↑
MFF-ESC	**20.64**	**8.87**	**4.81**	**2.98**	**18.83**
- w/ MFU v0	19.88	8.04	4.07	2.40	17.95
- w/ MFU v1	20.10	8.04	4.21	2.50	17.66
- w/ MFU v2	20.23	7.96	4.09	2.47	17.98
- w/ MFU v3	20.14	8.35	4.45	2.69	18.29

Table 4: Comparisons with variants of MFU.

As shown in Table 3, MFF-ESC without emotion intensity or feedback information results in the decline of **B-n** and **R-L**, indicating the effectiveness of multi-stream fusion. The significant decrease resulted by KL loss elimination verifies the effectiveness of using feedback to guide the learning of emotion intensity in latent space. Moreover, compared with merely removing a certain loss, most of the results drop more significantly when eliminating the corresponding information entirely. It further proves the modelling processes of emotion intensity, feedback, and problem type are effective and contribute to supportive response generation.

7.3. Comparison with Variants of MFU

As shown in Table 4, to evaluate the effect of the proposed MFU, we compare it with some variants of MFU. All the variants have utilized the three streams (text semantics stream, emotion intensity stream, and feedback stream) as input. **MFU v0** means using two LSTM units to incorporate emotion intensity stream and feedback stream, respectively. Besides, the architecture of LSTM hasn't been changed. Semantics-emotion intensity fusion and semantics-feedback fusion are implemented at the input stage of LSTM unit. **MFU v1** means using one LSTM unit and fusing the three streams at the input stage, without revision of LSTM unit. **MFU v2** means using one LSTM unit and revising the output gate, concatenating emotion intensity and feedback embeddings at the output gate stage. **MFU v3** is similar to MFU that has two output gates as shown in Figure 3, the difference is **MFU v3** conducts multi-stream fusion at the input stage rather than injecting emotion intensity or feedback directly for the calculation of output gates.

Results reported in Table 4 indicate the superiority of our proposed MFU that uses feedback stream and emotion intensity stream to directly influence the calculation of output gate instead of affecting all the gates. In this way, the long-term memory of dialogue history and short-term memory of current feedback and emotion intensity can both be well maintained.

Situation	My mom does not understand me. I will say something and she thinks I am saying something else.
User	I am a little angry. My mother and I had another blow-up. *(predicted emotion intensity: **4**)*
TransESC	I am sorry to hear that. What's the problem?
MFF-ESC	**Oh no!** I am **so sorry to hear that**. What happened? (*Stronger Empathy*)
Ground-truth	Is your bit of anger sparked from the interaction with your mother?
	...
User	I try to tell her it isn't what I mean. It's like she doesn't understand. (*predicted emotion intensity: **4***)
TransESC	I can understand why you would feel that way. I would be too.
MFF-ESC	I can understand how frustrating that would be. **I've had to do the same thing with my mother. I would suggest you to have a conversation with her.** (*Self-disclosure & Providing Suggestions*)
Ground-truth	It is hard to feel misunderstood. Do you have some examples?
	...
User	Maybe I could just take deep breath and kind of step back a minute before I respond. (*predicted emotion intensity: **3***)
TransESC	Do you think that would help?
MFF-ESC	That sounds like **a great idea! I think that would be a great start.** (*Affirmation and Reassurance*)
Ground-truth	It would be good to do that. Maybe try to have a conversation and tell her you support her no matter what she's going through.
	...
User	I'll try to say things that don't sound confrontational from now on. (*predicted emotion intensity: **2***)
TransESC	I think that would be a great idea.
MFF-ESC	**I'm glad that I could help and always being here to support you.** (*Providing Support*)
Ground-truth	I believe in you.

Table 5: A case study of the transition of predicted emotion intensity.

	Feedback	**Init. Emo.**	**Fin. Emo.**
Accuracy (%)	83.76	93.75	26.34
	Declined	**Maintained**	**Increased**
p_1 **vs.** p_{init}	100.00%	0	0
p_2 **vs.** p_1	50.00%	10.82%	39.18%
p_{fin} **vs.** p_2	17.37%	62.55%	20.08%
p_{fin} **vs.** e_{init}	72.96%	26.32%	0.72%

Table 6: Transition of the predicted emotion intensity in MFF-ESC. **Init.** and **Fin.** are short for initial and final. p_{init} and p_{fin} mean the predicted emotion intensity at the beginning and end of the conversation, respectively. p_1 and p_2 mean the predicted emotion intensity at the one-third position and two-thirds position of the dialogue, respectively. e_{init} means the true initial emotion intensity.

7.4. Transition of the Predicted Emotion Intensity

To evaluate the transition of the predicted emotion intensity, we calculate the accuracy of feedback score prediction and emotion intensity prediction during the testing process.

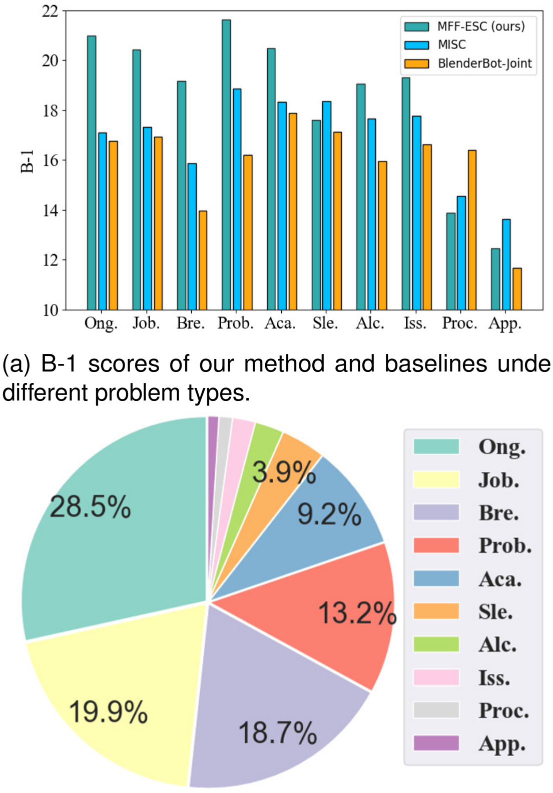
As shown in the upper part of Table 6, the accuracy of feedback prediction and initial emotion intensity prediction is much higher than that of final emotion intensity. To investigate the reason for it, we analyse the prediction of emotion intensity **at the beginning** (p_{init}), **the one-third position** (p_1), **the two-thirds position** (p_2), and **the end of** (p_{fin}) the conversation, evaluating whether the emotion intensity representations have learned the dynamic transition of user's emotion intensity.

As shown in the lower part of Table 6, **100%** of p_1 are lower than p_{init}. **50%** of p_2 are lower than p_1 while **39.18%** of p_2 are higher than p_1. **17.37%** of p_{fin} are lower than p_2 while **20.08%** of p_{fin} are higher than p_2. It indicates that the emotion intensity vectors in MFF-ESC have learned the transition of emotion intensity indeed but it's still difficult to make the predicted emotion intensities keep falling along with the dialogue.

However, when we only compare the true initial emotion intensity e_{init} with the predicted final emotion intensity p_{fin}, **72.96%** of p_{fin} achieve a decline. The results indicate that although the accuracy of final emotion intensity prediction isn't desirable, most of the predicted final emotion intensities are still lower than the initial intensities, proving the effectiveness of emotion intensity learning in MFF-ESC.

7.5. Case Study

We report a case in Table 5 to show how the modeling of emotion intensity affects the generated responses. At the beginning of the conversation, with the awareness of the high predicted emotion intensity, MFF-ESC generates a response with **stronger empathy** compared with TransESC. If the predicted emotion intensity is still high, MFF-ESC tries to **conduct self-disclosure or provide suggestions** to relieve the user's emotional distress.



(a) B-1 scores of our method and baselines under different problem types.

(b) Distribution of problem type on the test set.

Figure 4: Experiments of different problem type. From left to right of the abscissa in (a), the abbreviations represent *Ongoing Depression* (**Ong.**), *Job Crisis* (**Job.**), *Breakup with Partner* (**Bre.**), *Problems with Friends* (**Prob.**), *Academic Pressure* (**Aca.**), *Sleep Problems* (**Sle.**), *Alcohol Abuse* (**Alc.**), *Issues with Children* (**Iss.**), *Procrastination* (**Proc.**), and *Apperance Anxiety* (**App.**), respectively.

As the conversation goes, when the user finds out a suitable solution and the predicted emotion intensity declines, MFF-ESC **gives affirmation and reassurance or provides support**.

As shown in Table 5, the awareness of user's predicted emotion intensity during the conversation helps the model generate effective responses, leading to the relief of user's emotional distress.

7.6. Experiments about Problem Type

As shown in Figure 4(a), we compare our method with BlenderBot-Joint and MISC and report the results intuitively. From left to right of the abscissa, the proportion of problem type decreases. MFF-ESC performs worse than other methods on three problem types (Sle., Proc., and App.). However, utterances with these three problem types only account for **6.1%** on the test set. It is observed that MFF-ESC outperforms other baselines on most of the problem types which accounts for **93.9%** on the test set, proving the effectiveness of our method.

8. Conclusion

In this paper, we propose MFF-ESC which fuses three streams (text semantics stream, feedback stream, emotion intensity stream) thoroughly based on a subtly designed Multi-stream Fusion Unit, aiming to model the transition of emotion intensity. To further guide the learning of emotion intensity, we use KL divergence to minimize the distance between feedback distribution and emotion intensity distribution. Experimental results indicate the effectiveness of our method.

9. Limitations

The ESConv dataset contains long conversations with 29.8 utterances per dialogue on average. As a result, how to model the subtle transition of emotion intensity as the conversation goes is an intractable challenge. In this paper, we fully use the emotion intensity-related knowledge within the dataset and propose a Multi-stream Fusion Unit (MFU) to incorporate the three relative streams (text semantics stream, emotion intensity stream, and feedback stream). Moreover, to further guide the representation learning of emotion intensity, we utilize KL divergence to minimize the distance between feedback distribution and emotion intensity distribution. Although most of the predicted final emotion intensities are lower than the initial intensities, the accuracy of final emotion intensity prediction still leaves a lot of room for improvement. We would like this paper to arouse researchers' interest in the study of negative emotion intensity modelling for ESC. In the future, we will further explore the emotion intensity modelling for ESC and extend MFU to other tasks.

10. Ethics Statement

The ESConv dataset (Liu et al., 2021a) used in this paper is publicly available. In the process of dataset collection, the information about individual privacy has been eliminated. For the human evaluation process in this paper, the participants are aware that the usage of annotation is only for research and monetary rewards are provided.

Furthermore, ESC aims to provide support through social interactions, such as the support from peers, friends, or families, rather than professional counseling (Liu et al., 2021a). In this paper, we don't claim to construct a chatbot for professional counseling which needs further efforts to guarantee the safety of generated responses, especially for serving users who tend to self-harm or suicide.

11. References

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

Brant R. Burleson, Meina Liu, Yan Liu, and Steven T. Mortenson. 2006. Chinese evaluations of emotional support skills, goals, and behaviors. *Commun. Res.*, 33(1):38–63.

Mao Yan Chen, Siheng Li, and Yujiu Yang. 2022. Emphi: Generating empathetic responses with human-like intents. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1063–1074. Association for Computational Linguistics.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 535–554. Association for Computational Linguistics.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3014–3026. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4040–4054. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: general language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 320–335. Association for Computational Linguistics.

Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health behavior and health education: Theory, research, and practice*, 4:189–210.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

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

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

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021a. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers),* *Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020a. MIME: mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8968–8979. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020b. MIME: mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8968–8979. Association for Computational Linguistics.

Harald A Mieg. 2001. *The social psychology of expertise: Case studies in research, professional domains, and expert roles*. Psychology Press.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411.

Saif M. Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20, 000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 174–184. Association for Computational Linguistics.

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

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. FADO: feedback-aware double controlling network for emotional support conversation. *Knowl. Based Syst.*, 264:110340.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy.

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

Petr Slovák, Ran Gilad-Bachrach, and Geraldine Fitzpatrick. 2015. Designing social and emotional skills training: The challenges and opportunities for technology support. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 2797–2800. ACM.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 308–319. Association for Computational Linguistics.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. 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, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023a. Is chatgpt equipped with emotional dialogue capabilities? *CoRR*, abs/2304.09582.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023b. Transesc: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6725–6739. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 813–824. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *CoRR*, abs/2308.11584.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6556–6566. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1714–1729. Association for Computational Linguistics.