Multi-stream Information Fusion Framework for Emotional Support Conversation

中國科學院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

LREC-COLING 2024

Reported by Yinan Bao

Outline

Introduction

Approach

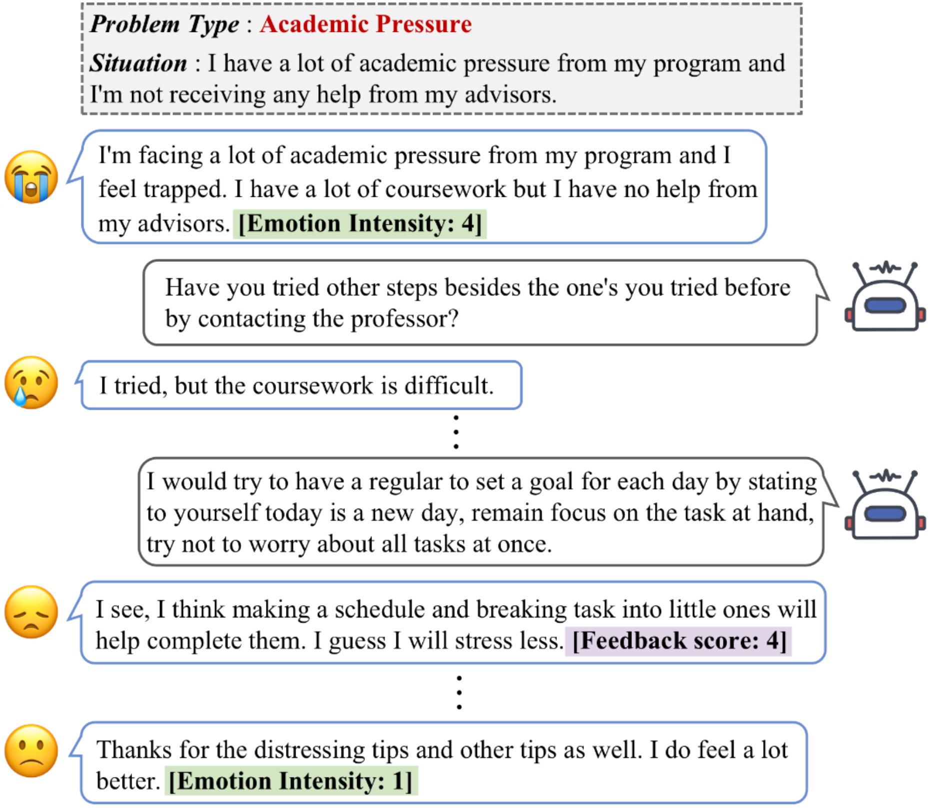
Experiments

Introduction

Emotional Support Conversation (ESC)

ESC is a task about response generation which aims to relieve the emotional distress of users who have high-intensity of negative emotions.

Unlike traditional emotional tasks, ESC system should perceive the transition of user's emotion intensity and generate supportive responses accordingly to decrease the intensity.



An example of emotional support conversation.

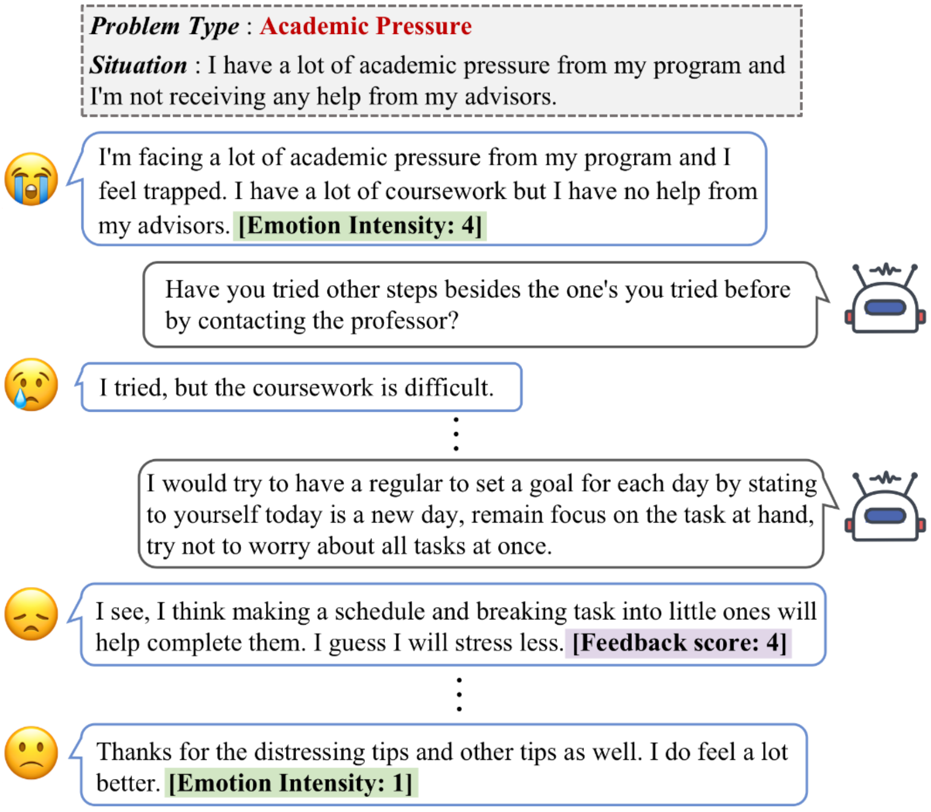
Introduction

1. Challenge

- **Effective Emotion Intensity Modeling:**

1. **The subtle change of emotion intensity is difficult to capture.**
2. **Each conversation only contains user's emotion intensity at the beginning and end of the conversation.**

How to model user's emotion intensity during the conversation.



An example of emotional support conversation.

2. Existing Methods

- **Overlook the modeling of emotion intensity.**

Introduction

3. Our method

a) 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).

b) 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.

c) Experiments on the benchmark dataset demonstrate the superiority of MFF-ESC, compared with state-of-the-art methods.

Approach

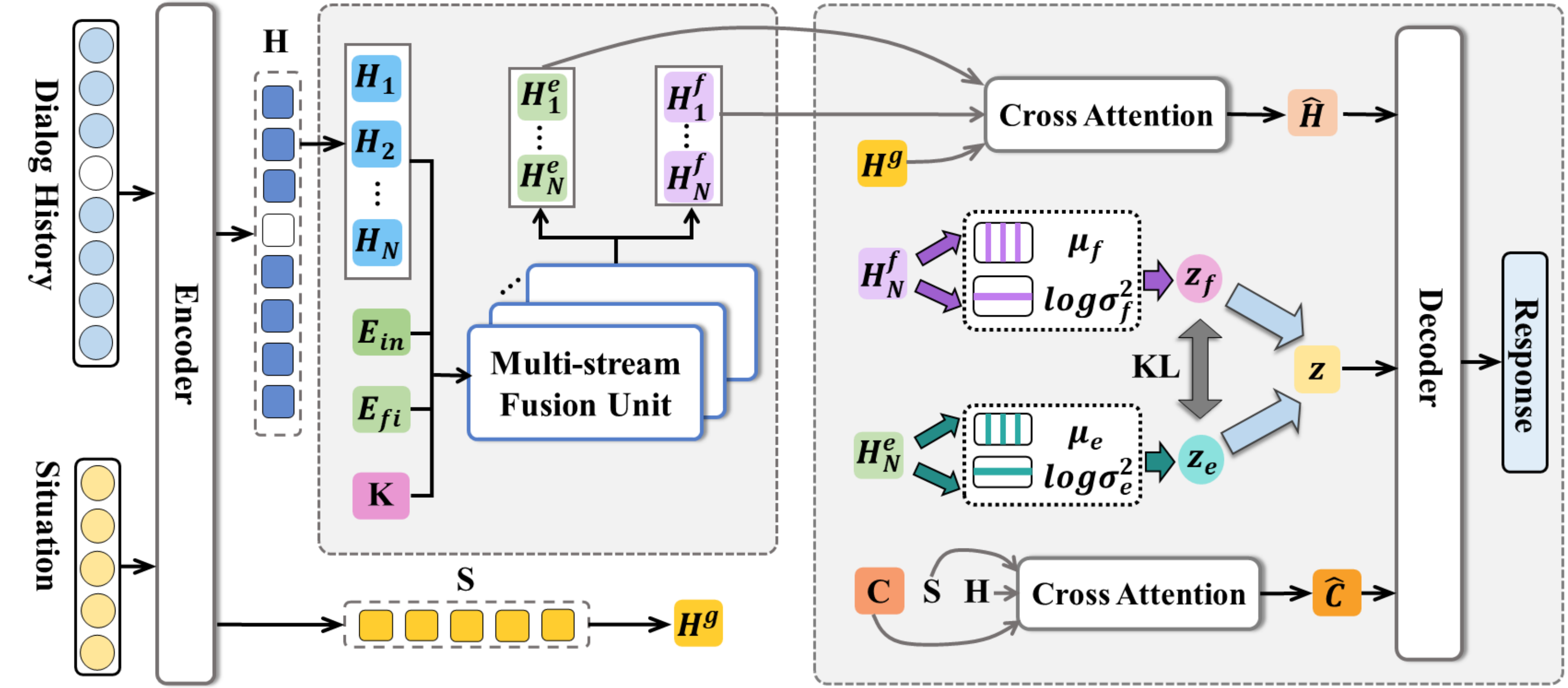
Problem Formulation

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 $\mathbf{p}(\mathbf{r}|D, \mathbf{e}_{in}, \mathbf{e}_{fi}, K, C, \mathbf{s})$.

[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.

Approach



Context Encoder & Multi-stream Fusion Unit

Response Generator

Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

Approach

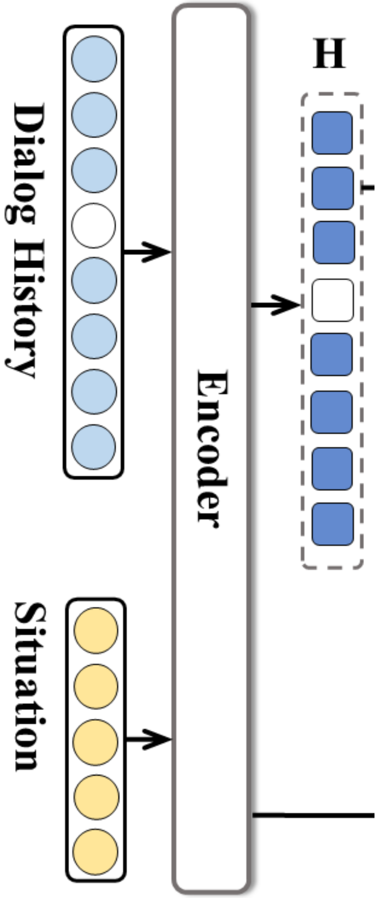
1. Context Encoder

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.

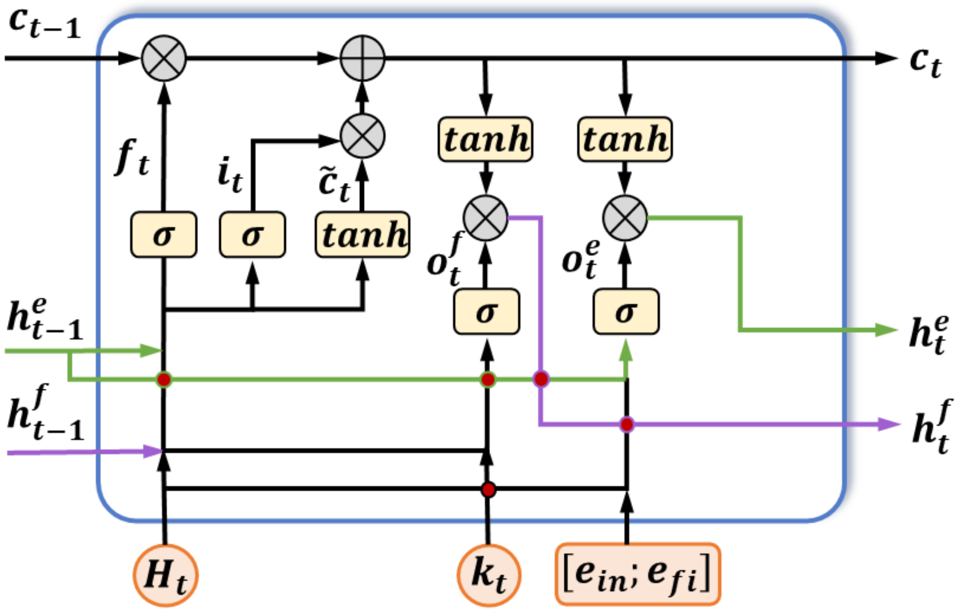


Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

Approach

2. Multi-stream Fusion Unit

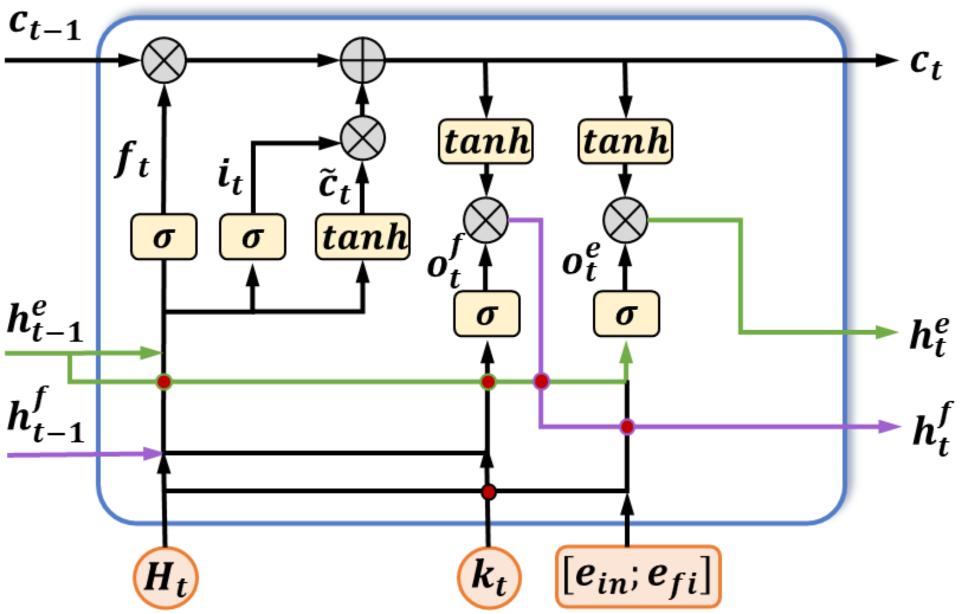
To model the transition of emotion intensity, we subtly design a Multi-stream Fusion Unit (MFU) based on LSTM to fuse three streams thoroughly, including text semantics stream, emotion intensity stream, and feedback stream.



Approach

2. Multi-stream Fusion Unit

- The difference between MFU and LSTM is **MFU has two output gates** for calculating the hidden states of emotion intensity and feedback respectively.



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

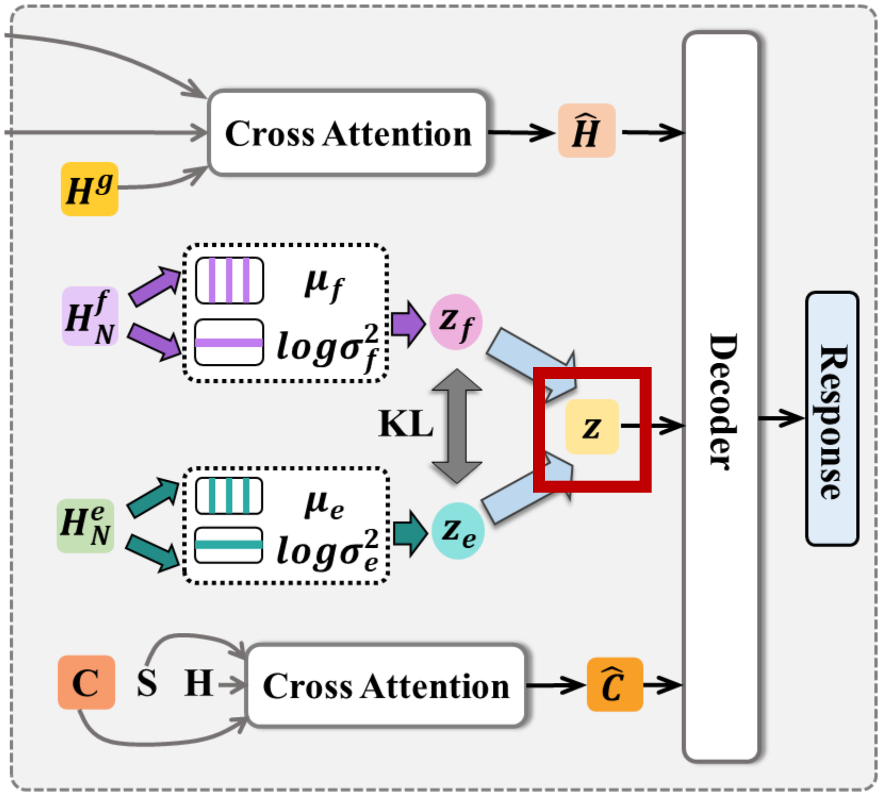
$$\mathbf{c}_t = \mathbf{f}_t * \mathbf{c}_{t-1} + \mathbf{i}_t * \tilde{\mathbf{c}}_t,$$
$$\mathbf{h}_t^e = \mathbf{o}_t^e * \mathtt{tanh}(\mathbf{c}_t),$$
$$\mathbf{h}_t^f = \mathbf{o}_t^f * \mathtt{tanh}(\mathbf{c}_t).$$

Approach

3. Response Generator

a. Injection 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 z from one of the two distributions for response generation.



Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

LREC-COLING 2024

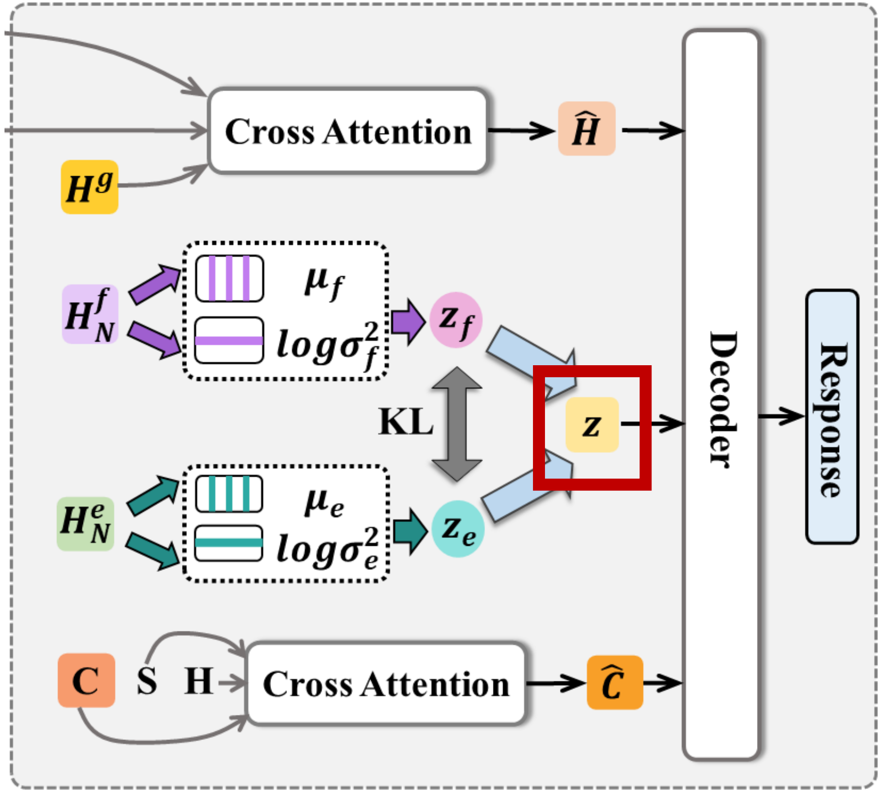
3. Response Generator

a. Injection of Emotion Intensity

- *Distribution of Emotion Intensity*

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:

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



Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

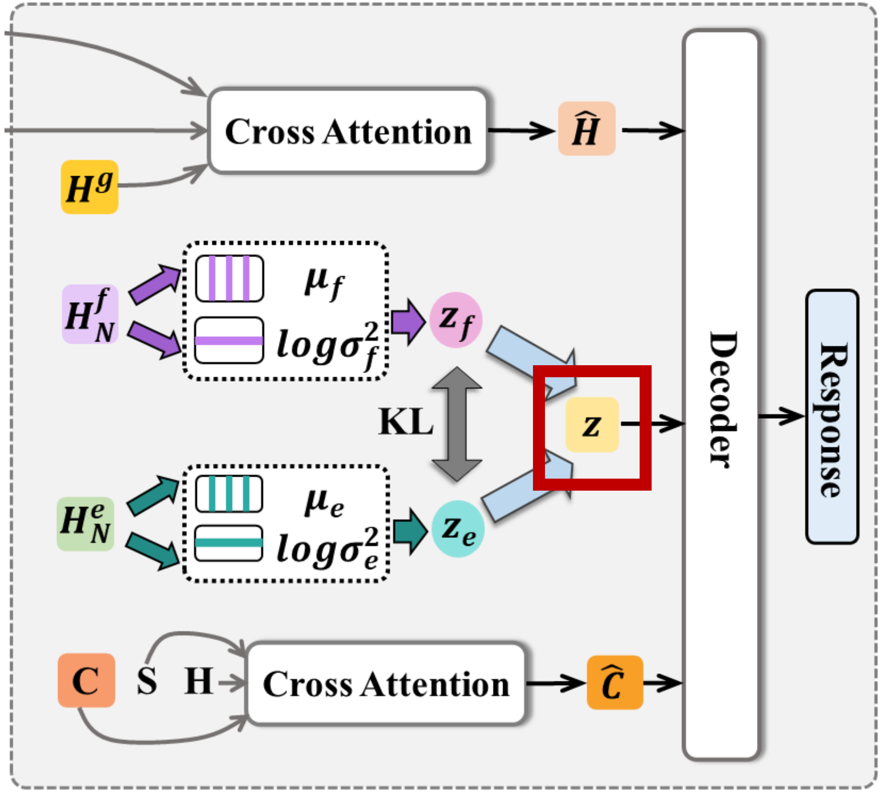
Approach

3. Response Generator

a. Injection of Emotion Intensity

- *Distribution of Feedback Score*

Similarly, we take the vector $\mathbf{H}_N^f$ of the last utterance as input, using a 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 $\mathbf{z}_f \in \mathbb{R}^{d_l}$.



Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

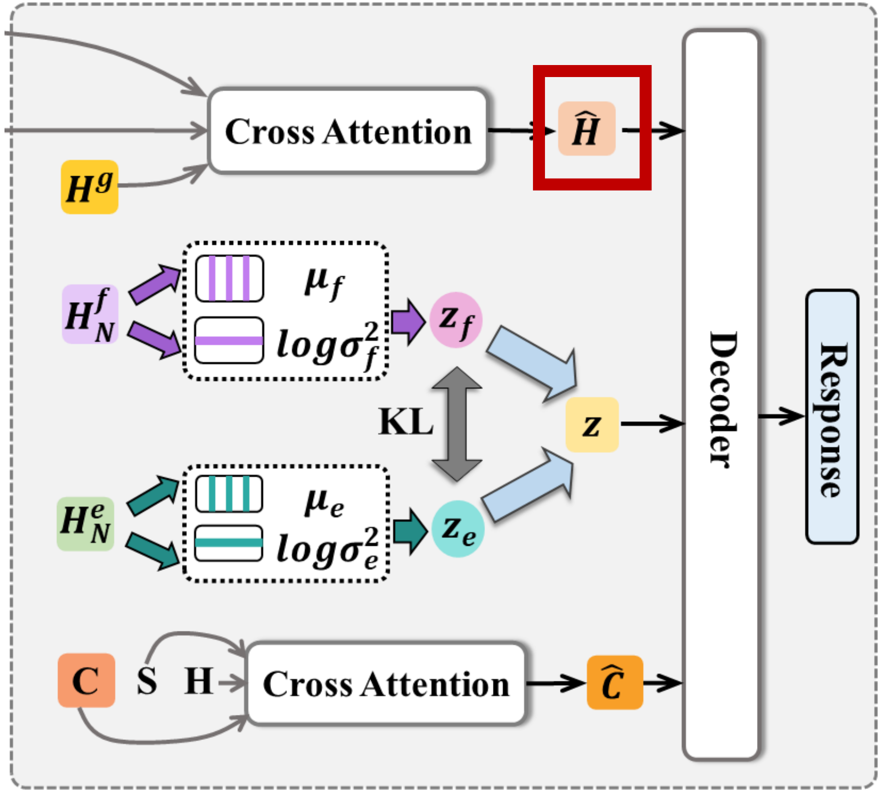
Approach

3. Response Generator

b. 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:

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



Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

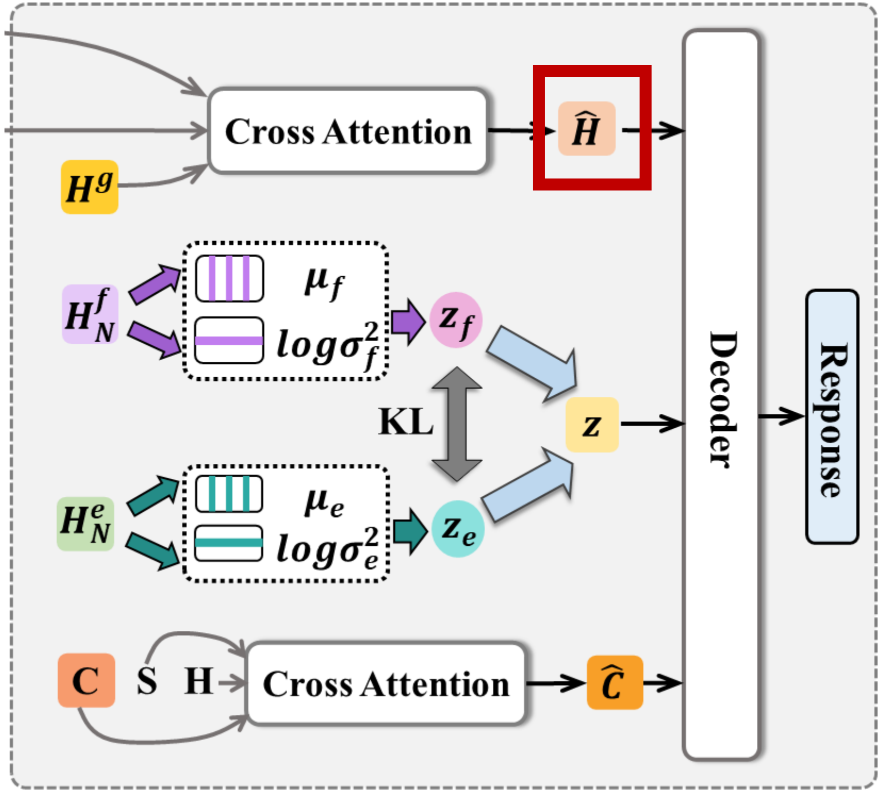
Approach

3. Response Generator

b. Injection of Situation

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

$$\begin{aligned}
\hat{\mathbf{H}}^g &= \mathrm{FNN}([\hat{\mathbf{H}}_e^g; \hat{\mathbf{H}}_f^g]), \\
\hat{\mathbf{H}}_N &= \mathrm{FNN}([\hat{\mathbf{H}}_N^e; \hat{\mathbf{H}}_N^f]), \\
\hat{\mathbf{H}} &= \tanh(\hat{\mathbf{H}}^g + \hat{\mathbf{H}}_N),
\end{aligned} \tag{6}$$



Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

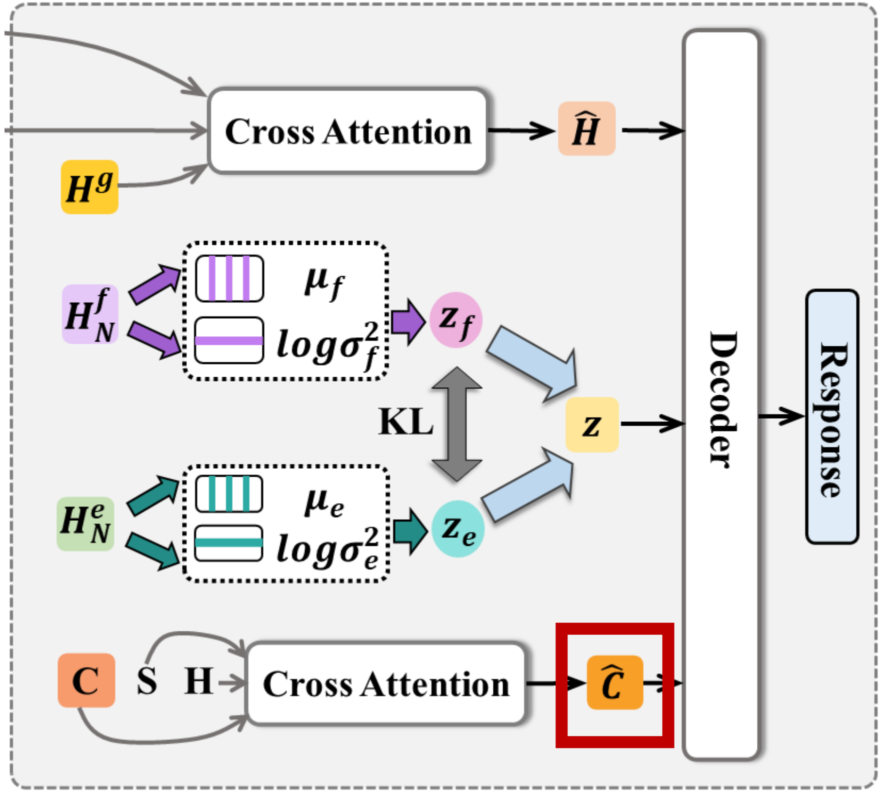
Approach

3. Response Generator

c. 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 $\mathbf{C}_{ex}$:

$$\boxed{\begin{aligned} \alpha_{ex} &= \mathbf{W}_2[\mathbf{C}\,;\,(\mathbf{S}\mathbf{W}_1+\mathbf{b}_1)]^\top+\mathbf{b}_2\,, \\ \alpha_{ex} &= \mathtt{softmax}(\alpha_{ex})\,, \\ \mathbf{C}_{ex} &= \mathbf{FNN}([\alpha_{ex}\mathbf{S}\,;\,\mathbf{C}])\,, \end{aligned}} \quad (7)$$



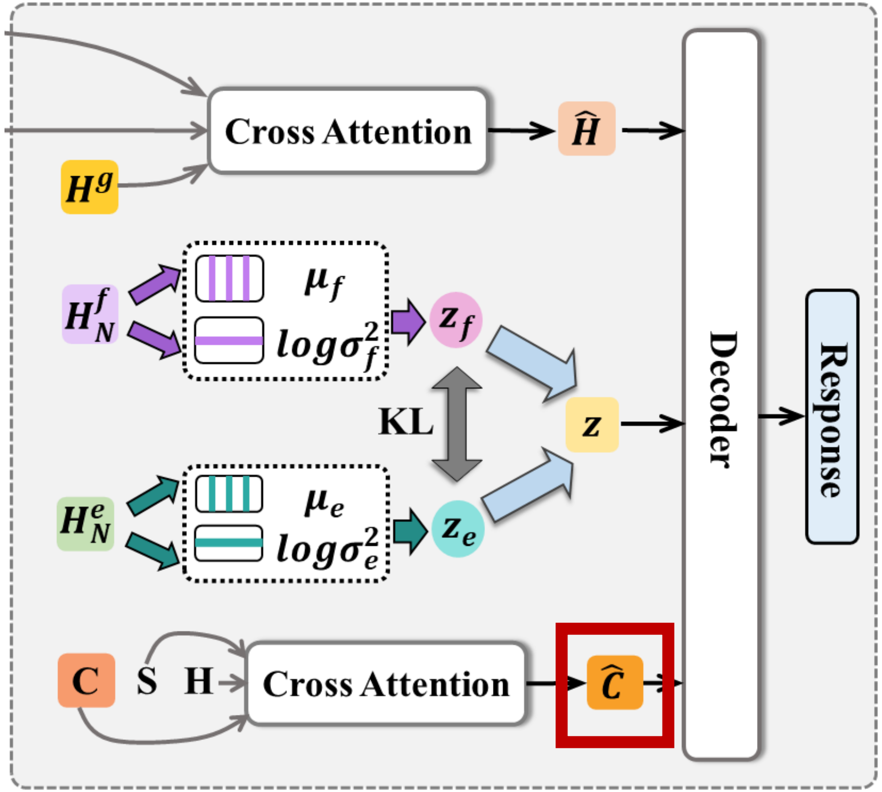
Approach

3. Response Generator

c. Injection 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:

$$\begin{aligned}
\alpha_{im} &= \mathbf{W}_4[\mathbf{C}\,;\,(\mathbf{H}\mathbf{W}_3 + \mathbf{b}_3)]^\top + \mathbf{b}_4\,, \\
\alpha_{im} &= \mathtt{softmax}(\mathbf{M} \odot \alpha_{im})\,, \\
\mathbf{C}_{im} &= \mathbf{FNN}([\alpha_{im}\mathbf{H}\,;\,\mathbf{C}])\,,
\end{aligned} \tag{8}$$



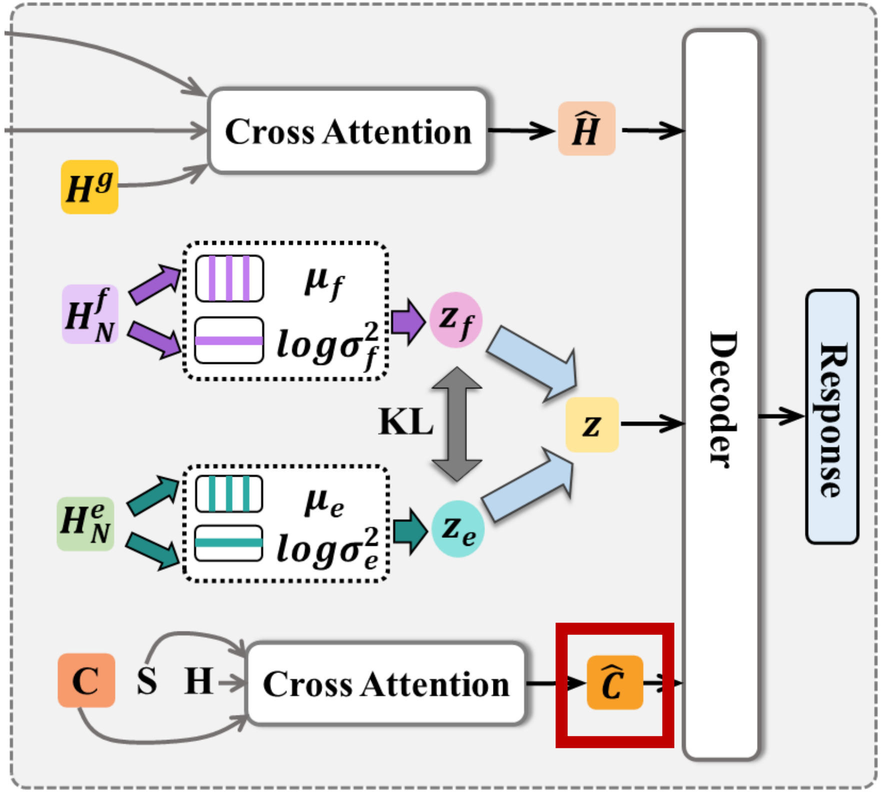
Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

3. Response Generator

c. Injection of Problem Type

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

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



Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

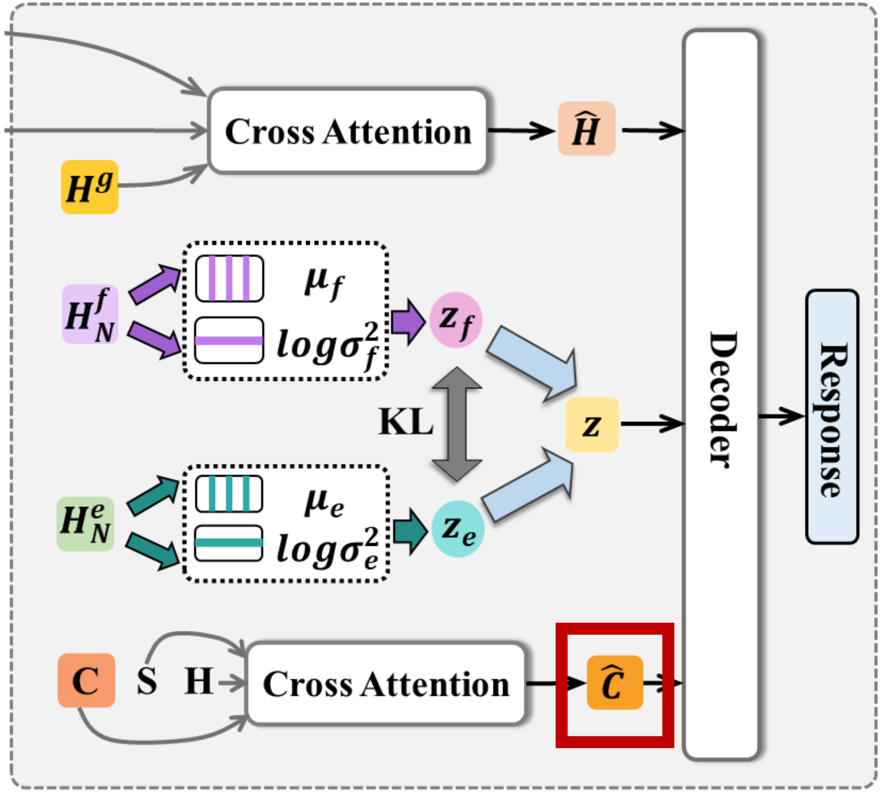
Approach

3. Response Generator

d. Response Generation

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:

$$\begin{aligned} \mathbf{O} &= \mathrm{FNN}([\hat{\mathbf{H}}\,;\,\mathbf{z}\,;\,\hat{\mathbf{C}}])\,, \\ \alpha &= \mathtt{softmax}(\mathbf{W}_5[\mathbf{O}\odot\mathbf{H}]^\top + \mathbf{b}_5)\,, \\ \mathbf{p}(\mathbf{r}|D,\mathbf{e}_{in},\mathbf{e}_{fi},K,C,\mathbf{s}) &= \mathbf{Dec}(\alpha\mathbf{O}+\mathbf{H})\,, \end{aligned} \tag{10}$$



Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

Approach

4. Loss Function

The reconstruction loss of a response with N_r tokens is:

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

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 $\mathcal{L}_f$.

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

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

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

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

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

Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

Experiments

1. Overall Results

- **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.**

Method	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	D-1↑	D-2↑	R-L↑
Transformer	81.55	17.25	5.66	2.32	1.31	1.25	7.29	14.68
MoEL	62.93	16.02	5.02	1.90	1.14	2.71	14.92	14.21
MIME	43.27	16.15	4.82	1.79	1.03	2.56	12.33	14.83
DialoGPT-Joint	19.41	17.06	6.22	2.87	1.57	2.82	17.30	15.03
BlenderBot-Joint	16.11	17.27	6.33	3.17	1.81	3.60	21.88	15.20
MISC	16.32	17.73	6.75	3.23	1.83	4.19	17.76	15.43
GLHG†	15.67	19.66	7.57	3.74	2.13	3.50	21.61	16.37
MultiESC‡	-	19.02	8.37	4.50	2.69	-	-	17.04
SUPPORTER	**15.39**	18.05	6.80	3.20	1.71	4.94	27.81	16.85
PAL	16.78	18.77	6.91	3.03	1.51	4.10	22.73	15.29
TransESC	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**

Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

Experiments

LREC-COLING 2024

2. Ablation Study

Eliminating the corresponding information completely in the modeling process.

Only removing the relative loss of the corresponding information.

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

Experiments

3. Comparison with Variants of MFU

a) **MFU v0** → Using two LSTM units to incorporate emotion intensity stream and feedback stream, respectively.

b) **MFU v1** → Using one LSTM unit and fusing the three streams at the input stage.

c) **MFU v2** → Using one LSTM unit and revising the output gate, concatenating emotion intensity and feedback embeddings at the output gate stage.

d) **MFU v3** → Conducting multi-stream fusion at the input stage rather than injecting emotion intensity or feedback directly for the calculation of output gates.

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

Experiments

4. Transition of the Predicted Emotion Intensity

- 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.

	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.

Experiments

5. Case Study

- As shown in the Table, 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.

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.

Multi-stream Information Fusion Framework for Emotional Support Conversation. (Bao et al., LREC-COLING 2024)

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

LREC-COLING 2024

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

Contact email: baoyinan@iie.ac.cn