MISC: A MIxed Strategy-Aware Model Integrating COMET for Emotional Support Conversation

Quan Tu[1*†], **Yanran Li**[2*], **Jianwei Cui**[2], **Bin Wang**[2], **Ji-Rong Wen**[1,3] and **Rui Yan**[1,3‡]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Xiaomi AI Lab
[3]Beijing Academy of Artificial Intelligence
[1]{quantu,jrwen,ruiyan}@ruc.edu.cn
[2]{liyanran,cuijianwei,wangbin11}@xiaomi.com

Abstract

Applying existing methods to emotional support conversation—which provides valuable assistance to people who are in need—has two major limitations: (a) they generally employ a conversation-level emotion label, which is too coarse-grained to capture user's instant mental state; (b) most of them focus on expressing empathy in the response(s) rather than gradually reducing user's distress. To address the problems, we propose a novel model **MISC**, which firstly infers the user's fine-grained emotional status, and then responds skillfully using a mixture of strategy. Experimental results on the benchmark dataset demonstrate the effectiveness of our method and reveal the benefits of fine-grained emotion understanding as well as mixed-up strategy modeling. Our code and data could be found in `https://github.com/morecry/MISC`.

1 Introduction

Empathy is the ability to perceive what others feel, think in their places and respond properly. It has a broad application scenarios to endow machines with the ability of empathy, including automatic psycho-therapist, intelligent customer service, empathetic conversational agents, and etc (Fitzpatrick et al., 2017; Shin et al., 2019; Ma et al., 2020).

In this work, we focus on a special kind of human-computer empathetic conversation, i.e., emotional support conversation (Liu et al., 2021). Distinguishedly, emotional support conversation happens between a seeker and supporter, where the supporter aims to gradually reduce seeker's distress as the conversation goes. This makes existing approaches unsuitable for our setting for at least two reasons. Firstly, existing work on emotional chatting learns to predict user emotion using a conversation-level emotion label, which is *coarse-grained* and *static* to the conversation context (Rashkin et al., 2019; Lin et al., 2019c; Li et al., 2020a). However, emotion is complex and user emotion intensity will *change* during the developing of the conversation (Liu et al., 2021). It is thus a necessity to tell seeker's *fine-grained* mental state at each utterance. Secondly, most of empathetic chatbots are trained to respond emotionally

---

*Equal Contribution.

†This work was done during internship at Xiaomi AI Lab.

‡Corresponding author: Rui Yan (ruiyan@ruc.edu.cn).

in accordance with the predicted coarse-grained emotion class, without consideration on how to address the seeker's emotional problem (De Graaf et al., 2012; Majumder et al., 2020; Xie and Park, 2021). Hence, they are deficient to apply for emotional support conversation whose goal is to help others work through the challenges they face.

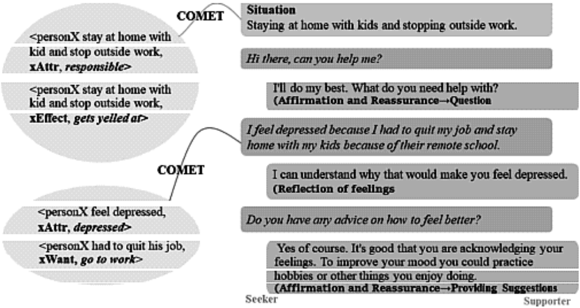


Figure 1: An Emotional Support Conversation Example.

To tackle these issues, we propose a novel approach **MISC**, a.k.a. **MI**xed **S**rategy-aware model integrating **COMET** for emotional support conversation. As to the first issue, we introduce COMET, a pre-trained generative commonsense reasoning model (Bosselut et al., 2019a), and devise an attention mechanism to selectively adopt the COMET knowledge tuples for fine-grained emotion understanding. As shown in Figure 1, this allows us to capture seeker's instantaneous mental state using different COMET tuples. In addition, we propose to also consider response strategy when generating empathetic responses for the second issue. Instead of modeling response strategy as a one-hot indi-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 308 - 319
May 22-27, 2022

cator, we formulate it as a probability distribution over a strategy codebook, and guide the response generation using a mixture of strategies. At last, our MISC produces supportive responses based on both COMET-enhanced mental information and distributed strategy representation. The unique design of mixed strategy not only helps to increase the expressed empathy, but also facilitates to learn the gradual transition in the long response, as the last utterance in Figure 1, which will in turn make the conversation more smooth.

To evaluate our model, we conduct extensive experiments on ESConv benchmark (Liu et al., 2021) and compare with 5 state-of-the-art empathetic chatbots. Based on both automatic metrics and manual judgments, we demonstrate that the responses generated by our model MISC are more relevant and empathetic. Besides, additional experimental analysis reveal the importance of response strategy modeling, and sheds light on how to learn a proper response strategy as well as how response strategy could influence the empathy of the chatbot.

In brief, our contributions are as follows: (1) We present a Seq2Seq model MISC, which incorporates commonsense knowledge and mixed response strategy into emotional support conversation; (2) We conduct experiments on ESConv dataset, and demonstrate the effectiveness of the proposed MISC by comparing with other SOTA methods. (3) We implement different ways of strategy modeling and give some hints on strategy-aware emotional support conversation.

2 Related Work

2.1 Emotion-aware Response Generation

As suggested in Liu et al. (2021), emotion-aware dialogue systems can be categorized into three classes: emotional chatting, empathetic responding and emotional support conversation. Early work target at emotional chatting and rely on emotional signals (Li et al., 2017; Zhou et al., 2018a; Wei et al., 2019; Zhou and Wang, 2018; Song et al., 2019). Later, some researchers shift focus towards eliciting user's specific emotion (Lubis et al., 2018; Li et al., 2020b). Recent work begin to incorporate extra information for deeper emotion understanding and empathetic responding (Lin et al., 2020; Li et al., 2020a; Roller et al., 2021). Li et al. (2021a) and Zhong et al. (2021) exploit ConceptNet to enhance emotion reasoning for response generation. Different from them, our work exploits a generative commonsense model COMET (Bosselut et al., 2019b), which enables us to capture seeker's mental states and facilitates strategy prediction in emotional support conversation.

2.2 Commonsense Knowledge for NLP

Recently, there is a large body of literature injecting commonsense knowledge into various NLP tasks, including classification (Chen et al., 2019; Paul and Frank, 2019), question answering (Mihaylov and Frank, 2018; Bauer et al., 2018; Lin et al., 2019a), story and language generation (Guan et al., 2019; Ji et al., 2020), and also dialogue systems (Zhou et al., 2018b; Zhang et al., 2020; Li et al., 2021a; Zhong et al., 2021). These dialogue systems often utilize ConceptNet (Speer et al., 2017), aiming to complement conversation utterances with physical knowledge. Distinguished from ConceptNet, ATOMIC (Sap et al., 2019) covers social knowledge including event-centered causes and effects as well as person-related mental states. To this end, ATOMIC is expected beneficial for emotion understanding and contributing to response empathy. In this work, we leverage COMET (Bosselut et al., 2019b), a commonsense reasoning model trained over ATOMIC for emotional support conversation.

2.3 Strategy-aware Conversation Modeling

Conversation strategy can be defined using different notions from different perspectives. A majority of research works is conducted under the notion of dialog acts, where a plethora of dialog act schemes have been created (Mezza et al., 2018; Paul et al., 2019; Yu and Yu, 2021). Dialog acts are empirically validated beneficial in both task-oriented dialogue systems and open-domain social chatbots (Zhao et al., 2017; Xu et al., 2018; Peng et al., 2020; Li et al., 2020c). As to empathetic dialogues, conversation strategy is often defined using the notion of response intention or communication strategy, which is inspired from the theories of empathy in psychology and neuroscience (Lubis et al., 2019; Li et al., 2021b). Whereas Welivita and Pu (2020) define a taxonomy of 15 response intentions through which humans empathize with others, Liu et al. (2021) define a set of 8 support strategies that humans utilize to reduce other's emotional distress. This partially reveals that response strategy is complex, which motivates us to condition on a mixture of strategy when generating supportive responses.

3 Preliminaries

3.1 ESConv Dataset

In this paper, we use the **E**motional **S**upport **Conv**ersation dataset, **ESConv** (Liu et al., 2021). Before conversations start, seekers should determine their emotion types, and tell the situation they are dealing with to supporters. Besides, the strategy of every supporter's utterance is marked, which is the most important to our work. In total, there are 8 kinds of strategies, and they are almost evenly distributed. More details are given in Appendix.

3.2 Problem Formulation

For general dialogue response generation, the target is to estimate the probability distribution $p(\boldsymbol{r}|\boldsymbol{c})$ of the dataset $\mathcal{D} = \{\boldsymbol{c}^{(i)}, \boldsymbol{r}^{(i)}\}_{i=1}^{N}$, where $\boldsymbol{c}^{(i)} = (\boldsymbol{u}_1^{(i)}, \boldsymbol{u}_2^{(i)}, ..., \boldsymbol{u}_{n_i}^{(i)})$ consists of a sequence of n_i utterances in the dialogue history, and $\boldsymbol{r}^{(i)}$ is the target response. For the sake of brevity, we omit the superscript (i) when denoting a single example in the remaining part.

In the setting of emotional support conversation, the seeker's situation $\boldsymbol{s}$ is considered as an extra input, which describes the seeker's problem in free-form text. We also denote the seeker's last post (utterance) as $\boldsymbol{x}$. Consequently, the target becomes to estimate the probability distribution $p(\boldsymbol{r}|\boldsymbol{c}, \boldsymbol{s}, \boldsymbol{x})$.

4 Model: MISC

The overview of our approach is shown in Figure 2. Based on blenderbot-small (Roller et al., 2021), our model MISC consists three main components: (1) a mental state-enhanced encoder (Bosselut et al., 2019a); (2) a mixed strategy learning module; and (3) a multi-factor-aware decoder.

4.1 Mental State-Enhanced Encoder

Following common practice, we firstly represent the context using the encoder $\mathtt{E}$:

$$\boldsymbol{C} = \mathtt{E}(\mathtt{CLS}, \boldsymbol{u}_1, \mathtt{EOS}, \boldsymbol{u}_2, ..., \boldsymbol{u}_{n_i}) \tag{1}$$

where CLS is the start-token and EOS is the separation-token between two utterances.

To better understand the seeker's situation, we exploit COMET (Bosselut et al., 2019a), a commonsense knowledge generator to supply mental state information related to the conversation. Concretely, we treat the situation $\boldsymbol{s}$ as an event, and feed it with different relations into COMET:

$$\boldsymbol{B}^s = \bigcup_{j=1}^{N_r} \texttt{COMET}(rel_j, \boldsymbol{s}) \tag{2}$$

where N_r is the number of pre-defined relations in COMET, and rel_j stands for the j-th specific relation, such as `xAttr` and `xReact`.[1] Note that given a certain event-relation pair, COMET is able to generate multiple "tails" of free-form mental state information, $\boldsymbol{B}^s$ is a set of N_s mental state blocks, i.e., $\boldsymbol{B}^s = \{\boldsymbol{b}_j^s\}_{j=1}^{N_s}$. Similarly, we can obtain the set of mental state blocks $\boldsymbol{B}^x$ using the seeker's last post $\boldsymbol{x}$.

Then, all of the free-form blocks will be transformed into dense vectors using our encoder E:

$$\begin{aligned} \hat{\boldsymbol{H}}^s &= [\boldsymbol{h}_{1,1}^s, \boldsymbol{h}_{2,1}^s, ..., \boldsymbol{h}_{N_{st},1}^s] \\ \boldsymbol{h}_j^s &= \mathtt{E}(\boldsymbol{b}_j^s) \end{aligned} \tag{3}$$

and the hidden state of each block's first token will be used to represent the corresponding block. Later, due to the noisy of COMET blocks, a lot of them are irrelevant to the context. We creatively take attention method to refine the strongly relevant blocks. That operation could be expressed as

$$\begin{aligned} \boldsymbol{Z} &= \mathtt{softmax}(\hat{\boldsymbol{H}}^s \cdot \boldsymbol{C}^{\mathrm{T}}) \cdot \boldsymbol{C} \\ \boldsymbol{H}^s &= \mathtt{LN}(\hat{\boldsymbol{H}}^s + \boldsymbol{Z}) \end{aligned} \tag{4}$$

where LN is the LayerNorm module (Ba et al., 2016). Similarly, we could transform $\boldsymbol{x}$ to $\boldsymbol{H}^x$ following the same method as $\boldsymbol{s}$ to $\boldsymbol{H}^s$. At last, we get the conversation-level and utterance-level representation of seeker's mental state $\boldsymbol{H}^s$ and $\boldsymbol{H}^x$, which are enhanced with commonsense information.

4.2 Mixed Strategy Learning Module

One straightforward way to predict the response strategy is to train a classifier upon the CLS states of the context representation $\boldsymbol{C}$ from Eq. (1):

$$\boldsymbol{p}^g = \texttt{MLP}(\boldsymbol{C}_1) \tag{5}$$

where MLP is a multi-layer perceptron, and $\boldsymbol{p}^g$ records the probabilities of each strategy to be used.

To model the complexity of response strategy as discussed before, we propose to employ the distribution $\boldsymbol{p}^g$ and model a mixture of strategies for

[1]Please refer to the appendix file for the definitions of all the relations as well as a brief introduction of COMET.

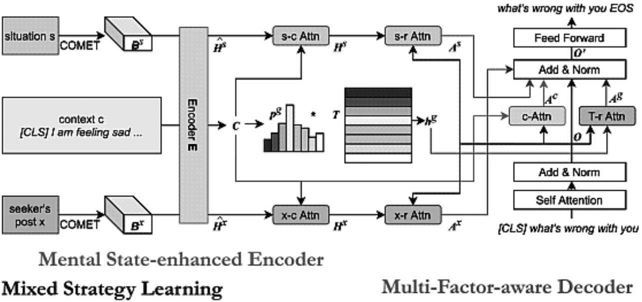


Figure 2: The overview of the proposed MISC which consists of a mental state-enhanced encoder, a mixed

strategy
learning module, and a multi-factor-aware decoder.

response generation. Here, we masterly learn from the idea of VQ-VAE's codebook to represent strategy(Oord et al., 2017). The strategy codebook $\boldsymbol{T} \in \mathbb{R}^{m \times d}$ represent m strategy latent vectors (here $m = 8$) with the dimension size d. By weighting $\boldsymbol{T}$ using $\boldsymbol{p}^g$, we are able to obtain a comprehensive strategy representation $\boldsymbol{h}^g$

$$\boldsymbol{h}^g = \boldsymbol{p}^g \cdot \boldsymbol{T} \tag{6}$$

Our codebook-based method has two benefits: (1) It is beneficial when long responses are needed to skillfully reduce the seeker's distress, which is common in emotional support conversation. (2) It is flexible to learn. Intuitively, if a strategy has a higher probability in $\boldsymbol{p}^g$, it should take greater effect in guiding the support conversation. In the extreme case where we have a sharp distribution, one single strategy will take over the control.

4.3 Multi-Factor-Aware Decoder

The remaining is to properly utilize the inferred mental states and the strategy representation. To notify the decoder of these information, we modify the backbone's cross attention module as:

$$
\begin{aligned}
\boldsymbol{A}^c &= \texttt{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{H}) \\
\boldsymbol{A}^s &= \texttt{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{H}^s) \\
\boldsymbol{A}^x &= \texttt{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{H}^x) \\
\boldsymbol{A}^g &= \texttt{CROSS-ATT}(\boldsymbol{O}, \boldsymbol{h}^g) \\
\boldsymbol{O}^{'} &= \texttt{LN}(\boldsymbol{A}^c + \boldsymbol{A}^s + \boldsymbol{A}^x + \boldsymbol{A}^g + \boldsymbol{O})
\end{aligned}
\tag{7}
$$

where CROSS-ATT stands for the backbone's cross attention module, and $\boldsymbol{O}$ is the hidden states of the decoder, which produces the final response by interacting with multi-factors.

Based on blenderbor-small (Roller et al., 2021), we jointly train the model to predict the strategy and produce the response:

$$
\begin{aligned}
\mathcal{L}_r &= -\sum_{t=1}^{n_r} \texttt{log}(p(r_t|\boldsymbol{r}_{j<t}, \boldsymbol{c}, \boldsymbol{s}, \boldsymbol{x})) \\
\mathcal{L}_g &= -\texttt{log}(p(g|\boldsymbol{c}, \boldsymbol{s}, \boldsymbol{x})) \\
\mathcal{L} &= \mathcal{L}_r + \mathcal{L}_g
\end{aligned}
\tag{8}
$$

where n_r is the length of response, g is the true strategy label, $\mathcal{L}_g$ is the loss of predicting strategy, $\mathcal{L}_r$ is the loss of predicting response, and $\mathcal{L}$ is combined objective to minimize.

5 Experiments

5.1 Experimental Setups

We evaluate our and the compared approaches on the dataset ESConv (Liu et al., 2021). For preprocessing, we truncate the conversation examples every 10 utterances, and randomly spilt the dataset into train, valid, test with the ratio of 8:1:1. The statistics is given in Table 1.

Category	Train	Dev	Test

# dialogues	14117	1764	1764
Avg. # words per utterance	17.25	17.09	17.11
Avg. # turns per dialogue	7.61	7.58	7.49
Avg. # words per dialogue	148.46	146.66	145.17

Table 1: The statistics of processed ESConv dataset.

5.2 Evaluation Metrics

We adopt a set of automatic and human evaluation metrics to assess the model performances:
Automatic Metrics. (1) We take the strategy prediction accuracy **ACC.** as an essential metric. A higher ACC. indicates that the model has a better capability to choose the response strategy. (2) We then acquire the conventional **PPL** (perplexity), **B-2** (BLEU-2), **B-4** (BLEU-4) (Papineni et al., 2002), **R-L** (ROUGE-L) (Lin, 2004) and **M** (Meteor) (Denkowski and Lavie, 2014) metrics to evaluate the lexical and semantic aspects of the generated responses. (3) For response diversity, we re-

port **D-1** (Distinct-1) and **D-2** (Distinct-2) numbers, which assesses the ratios of the unique n-grams in the generated responses (Li et al., 2016).

Human Judgments. Following See et al. (2019), we also recruit 3 professional annotators with linguistic and psychologist background and ask them to rate the generated responses according to Fluency, Knowledge and Empathy aspects with level of {0,1,2}. For fair comparison, the expert annotators do not know which model the response is from. Note that these 3 writers are paid and the results are proof-checked by 1 additional person.

5.3 Compared Models

Transformer is a vanilla Seq2Seq model trained based on the MLE loss (Vaswani et al., 2017).

MT Transformer is the **M**ulti-**T**ask transformer which considers emotion prediction as an extra learning task (Rashkin et al., 2018). In specific, we use the conversation-level emotion label provided in ESConv to learn emotion prediction.

MoEL softly combines the output states from mul-

tiple listeners (decoders) to enhance the response empathy for different emotions (Lin et al., 2019b).
MIME considers the polarity-based emotion clusters and emotional mimicry for empathetic response generation (Majumder et al., 2020).
BlenderBot-Joint is the SOTA model on ESConv dataset, which prepends a special strategy token before the response utterances (Liu et al., 2021).

5.4 Implementation Details

We implement our approach based on blenderbot-small (Roller et al., 2021) using the default sizes of vocabulary and the hidden states. For the last post $\boldsymbol{x}$ and the situation $\boldsymbol{s}$, we set the maximum number of the retrieved COMET blocks as 30 and 20 respectively. The inferred COMET blocks will be sent to the encoder with a maximum of 10 words.

To be comparable with the SOTA model in Liu et al. (2021), we fine-tune MISC based on the blenderbot-small with the size of 90M parameters by a Tesla-V100 GPU. The batch size of training and evaluating is 20 and 50, respectively. We initialize the learning rate as 2e-5 and change it during training using a linear warmup with 120 warmup steps. We use AdamW as optimizer (Loshchilov and Hutter, 2018) with β_1=0.9, β_2=0.999 and ϵ=1e-8. After training 8 epochs, the checkpoint with the lowest perplexity on the validation set is selected for testing. Following (Liu et al., 2021), we also adopt the decoding algorithms of Top-p and Top-k sampling with p=0.3, k=30, temperature τ=0.7 and the repetition penalty 1.03. We will release the source code to facilitate future work.

5.5 Experimental Results

As shown in Table 2, the vanilla Transformer performs the worst according to its relatively low PPL, BLEU-n and distinct-n scores. This is not suprising because it does not have any other specific optimization objective to learn the ability of empathy, and it is observed to be deficient for capturing long context as that in the ESConv dataset.

The performances of MT Transformer, MoEL and MIME, are also disappointing. Even though they three are equipped with empathetic objectives such as emotion prediction and ensembling listener, they are based on the conversation-level static emotion label, which is not adequate for fine-grained emotion understanding. More importantly, these three empathetic models lack of the ability of strategically consoling the seekers in the setting of emotional support conversation.

By comparing with the SOTA model BlenderBot-Joint, we can see that our model MISC is more effective especially in predicting more accurate response strategy. Whereas BlenderBot-Joint predicts one single strategy at the first decoding step, our method MISC models mixed response strategies using a strategy codebook and allows the decoder to learn the smooth transition and exhibit empathy more naturally. The comparison result suggests that it is beneficial to predict the response strategy as an extra task and to take into consideration the strategy complex for emotional support conversation.

The human evaluation results in Table 3 are consistent with the automatic results. Thanks to the pre-

Model	ACC(%)↑	PPL↓	D-1↑	D-2↑	B-2↑	B-4↑	R-L↑	M(%)↑
Transformer	-	89.61	1.29	6.91	6.53	1.37	15.17	10.33
MT Transformer	-	89.52	1.28	7.12	6.58	1.47	14.75	10.27
MoEL	-	133.13	2.33	15.26	5.93	1.22	14.65	9.75
MIME	-	47.51	2.11	10.94	5.23	1.17	14.74	9.49
BlenderBot-Joint	28.57	18.49	4.12	17.72	5.78	1.74	16.39	9.93
MISC	**31.63**	**16.16**	**4.41**	**19.71**	**7.31**	**2.20**	**17.91**	**11.05**

Table 2: Automatic Evaluation Results on ESConv.

Model	Flu.	Know.	Emp.
Transformer	0.62	0.31	0.29
MT Transformer	0.78	0.34	0.82
MoEL	0.36	0.80	0.33
MIME	1.13	0.27	0.35
BlenderBot-Joint	**1.87**	0.74	1.21
MISC	1.84	**1.06**	**1.44**

Table 3: Manual Evaluation Results. The Fleiss Kappa

score (Fleiss and Cohen, 1973) reaches 0.445, indicating a moderate level of agreements.

trained LM blenderbot-small (Rashkin et al., 2018), BlenderBot-Joint and our MISC significantly outperform other models on the Fluency aspect. Notably, our MISC yields the highest Knowledge score, which indicates that the responses produced by our approach contain much more specific information related to the context. We conjecture that our multi-factor-aware decoder successfully learns utilize the mental state knowledge from COMET with the mixture of the predicted strategies.

Overall speaking, MISC performs the best on almost every metric. It strongly demonstrates the effectiveness of our approach, and highlights the importance of fine-grained mental state modeling and mixed response strategy incorporation.

6 Analysis

Our method MISC has two novel designs: considering the fine-grained mental states and incorporating a mixture of response strategy. To investigate more, we conduct extra experiments, and the analysis results give us hints of how to develop better emotional support conversational agents.

6.1 Ablation Study

In order to verify the improvement brought by each added part (g, s, x), we drop these three parts from the MISC and check the performance changes. As shown in Table 4, the scores on all the metrics decrease dramatically when the g is albated. Consequently, we suppose the strategy attention is vital for guiding the semantics of the response. In addition, the scores also decline when we remove the the situation s and the seeker's last query x. According to the above experiments, each main part of the MISC is proven effective.

Model	D-1↑	B-2↑	R-L↑	M(%)↑
MISC	**4.41**	**7.31**	**17.91**	**11.05**
w/o g	3.85	7.09	16.75	9.85

w/o s	4.39	6.35	17.05	10.06
w/o x	4.27	6.49	17.03	10.09

Table 4: Evaluation Results of Ablation Study.

6.2 Case Study

In Table 5, an example is present to compare the response generated by the MISC and the other models. Various problems appear in the compared models, such as inconsistency, repetition, contradiction, etc. Intuitively, our model achieves the best performance in contrast. Besides, we present a visualization in Figure 4 to interpret how the MISC organizes the response under the combined effect of the COMET blocks and the mixture of strategies.

6.3 Fine-grained Emotion Understanding

As discussed before, one limitation of previous approaches is that they solely rely on a conversation-level emotion label, which is too coarse to guide the chatbot respond strategically and help the emotional conversation progress healthily. To remedy this issue, we exploit the commonsense knowledge generator COMET to supplement fine-grained information of seeker's mental state.

In order to fairly examine the effects of different emotional information, we discard the COMET blocks and implement a variant of our method MISE, a.k.a. **MI**xed-**S**rategy-aware model integrating **E**motion, where an extra emotion classification objective is added to the main architecture, as in Rashkin et al. (2018). Table 6 summarizes the comparison results between our full model MISC and its variant MISE. Obviously, all the metrics

Situation	
Seeker	My boyfriend and I recently broke up due to long-distance relationship and the impact 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 break up?
Response	
Transformer	I am so sorry that you are not able to get your friends and family. (*Inconsistency*)
MT Transformer	I have had a similar situation and I understand. It was the only way I was in similar situation, I have found that way to get out of my life. I have ... and i was in a similar situation. (*Repetition*)
MIME	I know it is hard, but it is always good to do things out for yourself. (*Poor Empathy*)
MoEL	I am so sorry that you are feeling better. I had a relationship with your boyfriend! (*Aggressiveness*)
BlenderBot-Joint	I have not but i have been through a lot of breakups... (*Contradiction*)
MISC	I have been through a breakup and I can tell you that it was very painful. I have had to move on from a relationship because of it. (*Strong Empathy*)
Ground-Truth	It sounds like you are really stressed over the breakup and you want to know how to move forward.

Table 5: Responses generated from MISC and other compared models. Some words are omitted due to space limit.

Component	D-1 ↑	B-2 ↑	R-L ↑	M(%) ↑
MISC	**4.41**	**7.31**	**17.91**	**11.05**
MISE	3.94	7.09	16.93	10.53

Table 6: Results of MISC with Different Emotions.

drop when replacing the fine-grained mental information with coarse-grained emotion label.

To depict the advantage of fine-grained mental state information, we visualize the attended COMET blocks of the example in Table 5. As shown in Figure 4, our chatbot MISC pays much attention of those inferred knowledge that are beneficial for fine-grained emotion understanding and strategy-aware empathetic responding.

More specifically, the attended COMET blocks (`xReact`, *hurt*) and (`xAttr`, *sad*) permit our chatbot MISC to utter the words "*it was painful*" which reflects its understanding of the seeker's feeling. Besides, note that the COMET blocks with white background are retrieved using the situation information s, and the grey ones are collected using the seeker's last post x. Despite of some overlapping, the white and grey attended blocks do contain distinct and crucial mental state knowledge. This partially validates that s and x is complementary to each other, and they two are useful information for emotional support conversation.

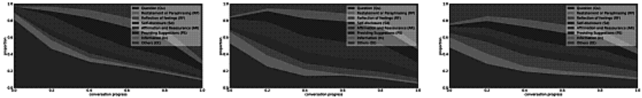
6.4 Mixed-Strategy-Aware Empathetic Responding

Meanwhile, the mixture of response strategy also plays a vital role for emotional support conversation. By analyzing the aforementioned case in depth, we find some hints on why our way to model conversation strategy is more preferred in the setting of emotional support conversation.

Hint 1: Mixed strategy is beneficial for Smooth Emotional Support. In Figure 4, we visualize the predicted strategy representation and the generated support response in Table 5. After understanding the seeker's situation of break-up and feelings of sadness, our MISC reasons that it might be proper to employ the strategies of *Self-disclosure*, *Reflection of feelings* to emotionally reply and effectively console the seeker's. Then, MISC organizes the response by firstly reveals that "it" has similar experiences and knows the feelings like. Moreover, the chatbot also supplements detailed information of *move on from a relationship* to suggest that the life will go on. These added-up words could be regarded as using the strategy of *Information* or *Others*, which is useful to transit the conversation to the next step smoothly. This case vividly shows how response generation is guided by the mixed strategies, and how skillful of our chatbot MISC is. **Hint 2: Mixed strategy is more effective than single strategy**. In addition to the case study, we also attempt to quantitatively assess the benefit of the mixed strategy modeling. To do so, we implement another variant of our chatbot **Single** where the mixed representation is replaced with an one-hot representation. Typically, we pick up the strategy dimension with the largest probability value as the one-hot output. The comparison results are given in

Table 7. Although yielding a slightly better distinct-n scores, the single-strategy variant lags far behind according to the lexical and semantic scores.

314



(a) predicted by the MISC.

(b) from the test set.

(c) predicted by the BlenderBot-Joint.

Figure 3: The strategy distribution in the different stage of conversation.

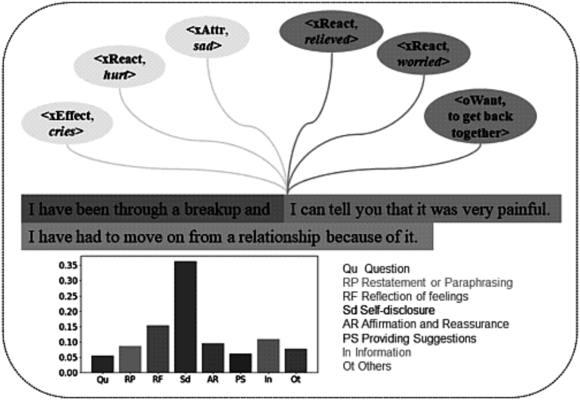


Figure 4: The visualization of how the MISC organizes the response under the effect of multiple factors.

Recall that the SOTA model BlenderBot-Joint (Liu et al., 2021) can also be regarded as a single-strategy model where a special strategy token is firstly decoded at the beginning of the response generation. We then compare their way of strategy modeling with our mixed strategy representation. As shown in Figure 5, the top-k strategy prediction accuracy of our MISC always surpasses that of BlenderBot-Joint, and the top-5 accuracy of our model reaches over 80%. This again proves the success of our strategy modeling.

Strategy	D-1↑	B-2↑	R-L↑	M(%)↑
Mixture	4.41	**7.31**	**17.91**	**11.05**
Single	**4.79**	6.30	17.01	10.22

Table 7: Comparison of different strategy modeling.

Hint 3: Mixed strategy is suitable for ESC Framework. The emotional support conversations in the dataset ESConv are guided by the ESC Framework, which suggests that emotional support generally follows a certain order of strategy flow. Similar to (Liu et al., 2021), here we also visualize the strategy distributions learned from different models, and compare them with the "ground-truth" strategy distribution in the original dataset. As shown in Figure 3, we can find: (1) Comparing our

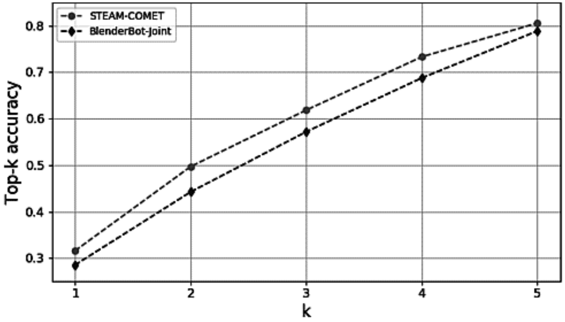


Figure 5: The Top-k Strategy Prediction Accuracy.

model with the SOTA model BlenderBot-Joint, we can find that our MISC better mimics the skill of strategy adoption in emotional support conversation. (2) At almost all stages of the conversation, our model is less likely to predict the strategy of *Others* (the grey part), as compared to BlenderBot-Joint. This indicates that the strategy acquired by our model is more discriminative than those by BlenderBot-Joint. (3) Overall speaking, the strategy distribution from our model share very similar patterns as compared to the ground-truth distribution. This implies that our way to model the strategy learning is suitable for the ESC framework.

7 Conclusions

In this paper, we propose MISC, a novel framework for emotional support conversation, which introduces COMET to capture user's instant mental state, and devises a mixed strategy-aware decoder to generate supportive response. Through extensive experiments, we prove the superiority and rationality of our model. In the future, we plan to learn the mixed response strategy in a dynamic way.

8 Ethical Considerations

At last, we discuss the potential ethic impacts of this work: (1) The ESConv dataset is a publicly-available, well-established benchmark for emotional support conversation; (2) **Privacy**: The original providers have filtered the sensitive information such as personally identifiable information (Liu et al., 2021); (3) Nevertheless, due to the limitation of filtering coverage, the conversations might still remain some languages that are emotionally triggering. Note that our work focuses on building emotional support conversational agents. For risky situations such as self-harm-related conversations, we do not claim any treatments or diagnosis.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive comments. This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62122089 & No. 61876196), Beijing Outstanding Young Scientist Program (NO. BJJWZYJH012019100020098), and Intelligent Social Governance Platform, Major Innovation Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. Rui Yan is the corresponding author, and is supported as a young fellow at Beijing Academy of Artificial Intelligence (BAAI).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019a. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019b. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Jiaao Chen, Jianshu Chen, and Zhou Yu. 2019. Incorporating structured commonsense knowledge in story completion. In *AAAI*.

Anneke De Graaf, Hans Hoeken, José Sanders, and Johannes WJ Beentjes. 2012. Identification as a mechanism of narrative persuasion. *Communication research*, 39(6):802–823.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Kathleen Kara Fitzpatrick, Alison M Darcy, and Molly Vierhile. 2017. 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 Mental Health*, 4.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613 – 619.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *AAAI*.

Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, Xiaoyan Zhu, and Minlie Huang. 2020. Language generation with multi-hop reasoning on commonsense knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 725–736, Online. Association for Computational Linguistics.

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

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020a. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. 2021a. Towards empathetic dialogue generation over multi-type knowledge. In *AAAI*.

Shifeng Li, Shi Feng, D. Wang, Kaisong Song, Y. Zhang, and Weichao Wang. 2020b. Emoelicitor: An open domain response generation model with user emotional reaction awareness. In *IJCAI*.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021b. Towards an online empathetic chatbot with emotion causes. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang

Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yanran Li, Ruixiang Zhang, Wenjie Li, and Ziqiang Cao. 2020c. Hierarchical prediction and adversarial learning for conditional response generation. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019a. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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)*, pages 2829–2839, Hong Kong, China. 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. 2019b. 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)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *AAAI*.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2019c. Caire: An empathetic neural chatbot. *arXiv preprint arXiv:1907.12108*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. 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 (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Nurul Lubis, S. Sakti, Koichiro Yoshino, and S. Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *AAAI*.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2019. Dialogue model and response generation for emotion improvement elicitation.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

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

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the*

27th International Conference on Computational Linguistics, pages 3539–3551, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Todor Mihaylov and Anette Frank. 2018. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 821–832, Melbourne, Australia. Association for Computational Linguistics.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

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*, pages 311–318.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3671–3681, Minneapolis, Minnesota. Association for Computational Linguistics.

Shachi Paul, Rahul Goel, and Dilek Z. Hakkani-Tür. 2019. Towards universal dialogue act tagging for task-oriented dialogues. *ArXiv*, abs/1907.03020.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online. Association for Computational Linguistics.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

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*, pages 300–325, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience lookahead. *arXiv preprint arXiv:1906.08487*.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695, Florence, Italy. Association for Computational Linguistics.

R. Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob

Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wei Wei, J. Liu, X. Mao, G. Guo, Feida Zhu, Pan Zhou, and Y. Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Baijun Xie and Chung Hyuk Park. 2021. Empathetic robot with transformer-based dialogue agent. In *2021 18th International Conference on Ubiquitous Robots (UR)*, pages 290–295. IEEE.

Can Xu, Wei Wu, and Yuehua Wu. 2018. Towards explainable and controllable open domain dialogue generation with dialogue acts. *ArXiv*, abs/1807.07255.

Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain HumanMachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1103–1120, Online. Association for Computational Linguistics.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2031–2043, Online. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 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 (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. CARE: commonsense-aware emotional response generation with latent concepts. In *AAAI*.

Hao Zhou, Minlie Huang, T. Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 1128–1137, Melbourne, Australia. Association for Computational Linguistics.

A Distribution of Strategies

As show in Figure 6, we can see that the proportion of each strategy is relatively balanced.

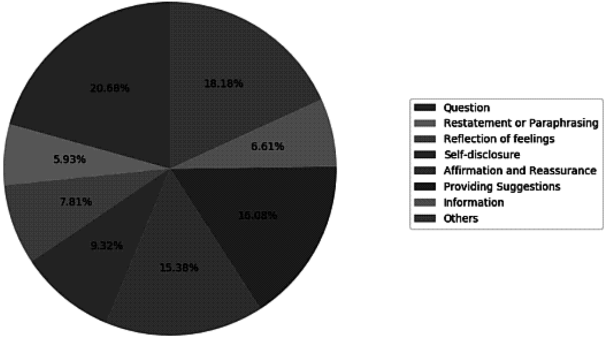


Figure 6: The strategy distribution in the original ES-Conv dataset.

B Definition of Strategies

Here, we directly adopted from (Liu et al., 2021) to help readers to learn about the specific meaning of each strategy more conveniently.

Question Asking for information related to the problem to help the help-seeker articulate the issues that they face. Open-ended questions are best, and closed questions can be used to get specific information.

Restatement or Paraphrasing A simple, more concise rephrasing of the help-seeker's statements that could help them see their situation more clearly.

Reflection of Feelings Articulate and describe the help-seeker's feelings.

Self-disclosure Divulge similar experiences that you have had or emotions that you share with the help-seeker to express your empathy.

Affirmation and Reassurance Affirm the help-seeker's strengths, motivation, and capabilities and provide reassurance and encouragement.

Providing Suggestions Provide suggestions about how to change, but be careful to not overstep and tell them what to do.

Information Provide useful information to the help-seeker, for example with data, facts, opinions, resources, or by answering questions.

Others Exchange pleasantries and use other support strategies that do not fall into the above categories.

C Description of COMET Relations

In the section, we also adopted the description from (Bosselut et al., 2019a), so as reader needn't to find it in original text.

`oEffect` The effect the event has on others besides Person X.

`oReact` The reaction of others besides Person X to the event.

`oWant` What others besides Person X may want to do after the event.

`xAttr` How Person X might be described given their part in the event.

`xEffect` The effect that the event would have on Person X.

`xIntent` The reason why X would cause the event.

`xNeed` What Person X might need to do before the event.
`xReact` The reaction that Person X would have to the event.
`xWant` What Person X may want to do after the event.