Knowledge-enhanced Mixed-initiative Dialogue System for Emotional Support Conversations

Yang Deng[1], **Wenxuan Zhang**[2,†], **Yifei Yuan**[1], **Wai Lam**[1]
[1] The Chinese University of Hong Kong, [2] DAMO Academy, Alibaba Group
{dengyang17dydy,isakzhang}@gmail.com
{yfyuan,wlam}@se.cuhk.edu.hk

Abstract

Unlike empathetic dialogues, the system in emotional support conversations (ESC) is expected to not only convey empathy for comforting the help-seeker, but also proactively assist in exploring and addressing their problems during the conversation. In this work, we study the problem of mixed-initiative ESC where the user and system can both take the initiative in leading the conversation. Specifically, we conduct a novel analysis on mixed-initiative ESC systems with a tailor-designed schema that divides utterances into different types with speaker roles and initiative types. Four emotional support metrics are proposed to evaluate the mixed-initiative interactions. The analysis reveals the necessity and challenges of building mixed-initiative ESC systems. In the light of this, we propose a knowledge-enhanced mixed-initiative framework (KEMI) for ESC, which retrieves actual case knowledge from a large-scale mental health knowledge graph for generating mixed-initiative responses. Experimental results on two ESC datasets show the superiority of KEMI in both content-preserving evaluation and mixed initiative related analyses.

1 Introduction

As the world is making efforts to recover from Covid-19 and plans for future construction, emotional support is of great importance in resolving the widespread emotional distress and increased risk for psychiatric illness associated with the pandemic (Pfefferbaum and North, 2020; Suh et al., 2021). A wide range of emotional support conversation (ESC) systems are emerging to provide prompt and convenient emotional support for help-seekers, including mental health support (Sharma et al., 2021; Lokala et al., 2022), counseling (Althoff et al., 2016; Shen et al., 2020, 2022) or motivational interviewing (Pérez-Rosas et al., 2016; Saha et al., 2021, 2022). Generally, the ESC system aims at reducing the user's emotional distress as well as assisting the user to identify and overcome the problem via conversations (Liu et al., 2021).

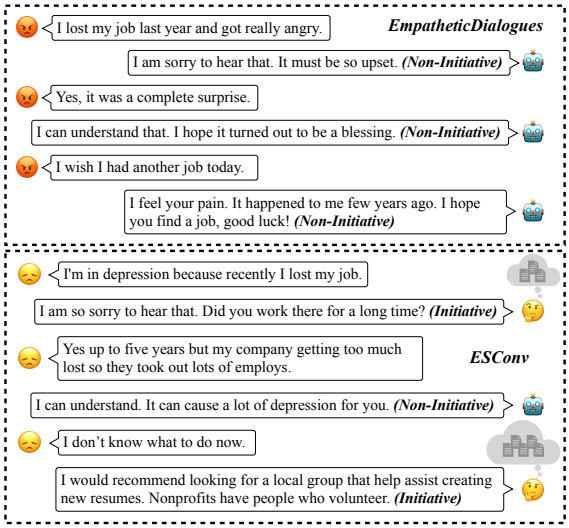


Figure 1: Examples from EMPATHETICDIALOGUES and ESConv datasets with a similar job loss problem.

Mixed initiative is commonly defined as an intrinsic feature of human-AI interactions where the user and the system can both take the initiative in leading the interaction directions (Allen et al., 1999; Kraus et al., 2020). For example, mixed-initiative conversational information-seeking (CIS) systems (Aliannejadi et al., 2019; Deng et al., 2023) can proactively initiate clarification interactions for resolving the ambiguity in the user query, instead of only reacting to the query. Accordingly, a mixed-initiative ESC system can proactively switch the initiative to provide an empathetic response or initiate a problem-solving discussion when appropriate. Many efforts have been made on the emotion reasoning for generating empathetic responses (Shen

* The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14200620).
† Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 4079–4095
July 9-14, 2023

et al., 2020; Zhang and Danescu-Niculescu-Mizil, 2020; Cheng et al., 2022; Peng et al., 2022). Another line of work focuses on identifying the dialogue acts of the utterances (Welivita and Pu, 2020; Malhotra et al., 2022; Svikhnushina et al., 2022) or predicting the next conversational strategies (Pérez-Rosas et al., 2017; Liu et al., 2021; Tu et al., 2022) in ESC systems. However, the feature of mixed initiative has not been investigated in existing ESC studies.

To facilitate the analysis on mixed-initiative ESC systems, we first propose an EAFR schema to annotate the utterances into different types with speaker roles and initiative types, named ***E**xpression* (User-initiative), ***A**ction* (Support-initiative), ***F**eedback* (User Non-initiative), and ***R**eflection* (System Non-initiative). Besides, four emotional support metrics are designed to measure the characteristics of initiative and non-initiative interactions in ESC, including *Proactivity*, *Information*, *Repetition*, and *Relaxation*.

To analyze the necessity of considering mixed initiative in ESC systems, we conduct a preliminary analysis on the different interaction patterns between ESC and empathetic dialogues (ED). Firstly, the dialogue flow analysis shows that the system in ED generally serves as a passive role, while the system in ESC proactively switches the initiative role during the conversation. As shown in Figure 1, the system in ED solely targets at comforting the user by reflecting their feelings or echoing their situations, *i.e.*, *Non-Initiative*. Differently, ESC systems are further expected to proactively explore the user's problem by asking clarifying questions and help the user overcome the problem by providing useful information or supportive suggestions, *i.e.*, *Initiative*. Furthermore, the analysis of the conversation progress and the emotional support metrics reveal three challenges in building a mixed-initiative ESC system: *1) When* should the system take the initiative during the conversation? *2) What* kind of information is required for the system to initiate a subdialogue? *3) How* could the system facilitate the mixed-initiative interactions?

According to these challenges, we define the problem of mixed-initiative ESC, which includes three sub-tasks: *1) Strategy Prediction* to determine the mixed-initiative strategy in the next turn, *2) Knowledge Selection* to collect the necessary knowledge for the next turn, and *3) Response Generation* to produce emotional support responses with appropriate mixed-initiative strategy and knowledge. To tackle this problem, we propose a novel framework, named Knowledge Enhanced Mixed-Initiative model (KEMI), to build a mixed-initiative dialogue system for emotional support conversations with external domain-specific knowledge. In detail, KEMI first employs a knowledge acquisition module to acquire emotional support knowledge from a large-scale knowledge graph on mental health dialogues. Specifically, we expand the user utterance with generated commonsense knowledge as a query graph and then perform subgraph retrieval over the knowledge graph. Secondly, a response generation module conducts multi-task learning of strategy prediction and response generation in a sequence-to-sequence manner to generate mixed-initiative responses with external knowledge.

The main contributions of this work are summarized as follows: (1) To measure the mixed-initiative interactions in ESC systems, we propose an innovative analysis method, including an EAFR annotation schema and corresponding emotional support metrics. (2) We propose a novel knowledge-enhanced mixed-initiative framework for ESC, which retrieves external knowledge from mental health knowledge graph by subgraph retrieval using the query graph expanded with commonsense knowledge. (3) Experimental results show that the mixed initiative is of great importance in ESC, and the proposed method effectively outperforms existing methods on both content-preserving evaluation and mixed initiative analyses.

2 Related Works

Emotional Support Conversation Similar to fine-grained sentiment analysis (Zhang et al., 2022, 2021c,b) in conversations (Li et al., 2022a; Zhang et al., 2021a), early works on emotional chatting mainly investigate approaches to detecting user emotions (Li et al., 2017; Zhou et al., 2018) or incorporating emotional signals into response generation (Wei et al., 2019; Song et al., 2019). As for empathetic dialogue systems (Rashkin et al., 2019; Welivita et al., 2021), evolving from emotion-aware response generation (Lin et al., 2019; Majumder et al., 2020) and emotional style transfer (Sharma et al., 2021), more efforts have been made on emotional reasoning techniques (Li et al., 2021; Kim et al., 2021; Gao et al., 2021; Cheng et al., 2022). Some latest studies explore the utilization of ex-

ternal knowledge for enhancing the model capability of emotion reasoning, including commonsense knowledge graph (Zhong et al., 2021; Li et al., 2022b), generative commonsense model (Sabour et al., 2021), and domain-specific knowledge (Shen et al., 2020, 2022). Shen et al. (2022) collectively exploit three kinds of external knowledge. Likewise, many ESC systems also leverage commonsense knowledge for response generation (Tu et al., 2022; Peng et al., 2022). However, the commonsense knowledge is rather abstractive without detailed information, so that it is less helpful for the ESC system to generate meaningful and informative responses. In this work, we employ the generative commonsense model for query expansion to retrieve actual case knowledge from an external knowledge graph.

Mixed-initiative Dialogue Recent years have witnessed many efforts on developing mixed-initiative conversational systems for various dialogues, such as information-seeking dialogues (Zamani et al., 2020; Aliannejadi et al., 2019), open-domain dialogues (Wu et al., 2019; Rachna et al., 2021; Lei et al., 2022), recommendation dialogues (Deng et al., 2021), conversational question answering (Deng et al., 2022a). Despite the importance of mixed initiative in ESC systems, this area has not been investigated. One closely related research scope is to recognize the conversation strategies (Liu et al., 2021; Pérez-Rosas et al., 2017) or the dialogue acts (Malhotra et al., 2022; Welivita and Pu, 2020; Svikhnushina et al., 2022; Deng et al., 2022b) of the utterances in ESC systems. However, these studies only focus on predicting the support strategies, instead of actually involving mixed-initiative interactions in ESC.

In addition, measuring mixed initiative is also regarded as an essential perspective for assessing dialogue quality (Vakulenko et al., 2021, 2020, 2019). Due to the high expenses in human evaluation, Sekulic et al. (2022) and Zhang and Balog (2020) investigate user simulation for evaluating the mixed-initiative interactions in conversational systems. In this work, we investigate several metrics for measuring the characteristics of the mixed initiative in ESC systems.

3 Preliminary Analysis

3.1 EAFR Schema & Metrics

Inspired by the ConversationShape (Vakulenko et al., 2021) for the analysis of mixed-initiative CIS systems, we first propose an EAFR annotation schema to study the mixed initiative in ESC systems. The EAFR annotation schema classifies the utterance in ESC into four categories w.r.t the role of speakers and the type of initiative, including ***Expression*** (User-initiative), ***Action*** (System-initiative), ***Feedback*** (User Non-Initiative), and ***Reflection*** (System Non-Initiative). Definitions and examples of each type are presented in Table 1.

Then, each utterance i in a dialogue is annotated as a tuple $(r_i, t_i, \boldsymbol{v}_i, e_i)$ for analysis. $r_i \in \{\text{User}(U), \text{System}(S)\}$ denotes the speaker role. $t_i \in \{\text{Initiative}(I), \text{Non-Initiative}(N)\}$ denotes the initiative type. $\boldsymbol{v}_i \in \{0,1\}^{|V|}$ denotes the one-hot vocabulary embeddings. $e_i \in [1,5]$ denotes the level of emotion intensity[1]. We further design four emotional support metrics for investigating patterns of mixed initiative in ESC systems as follows:

- **Proactivity**: how proactive is the system in the emotional support conversation?

$$\text{Pro} = \frac{1}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)} \sum_{i=1}^{n} \mathcal{I}(r_i = S, t_i = I) \quad (1)$$

  denotes the ratio of system-initiative interactions.

- **Information**: how much information does the system contribute to the dialogue?

$$\text{Inf} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{|V|} \mathcal{I}(r_i = S, v_{ik} = 1, \sum_{j=1}^{i-1} v_{jk} = 0)}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)} \quad (2)$$

  represents the average number of new frequent terms[2] that are introduced by the system.

- **Repetition**: how often does the system follow up on the topic introduced by the user?

$$\text{Rep} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{|V|} \mathcal{I}(r_i = S, v_{ik} = 1, \sum_{j=1}^{i-1} v_{jk}[r_j = U] > 0)}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)} \quad (3)$$

  represents the average number of repeated frequent terms that are introduced by the user and mentioned by the system.

- **Relaxation**: how well does the system relax the emotional intensity of the user?

$$\text{Rel}_i[r_i = S] = e_{<i}[r_{<i} = U] - e_{>i}[r_{>i} = U] \quad (4)$$

[1] A decrease from the intensity reflects emotion improvement (Liu et al., 2021).

[2] We only consider frequent terms that appear in the dialogue more than once. Standard pre-processing pipeline is adopted: remove punctuation, tokenization, lowercase, remove stopwords, and apply the English Snowball stemmer.

Role	Type	EAFR	Definition	Sample Utterances
User	Initiative	Expression	The user describes details or expresses feelings about the situation.	My school was closed due to the pandemic. I feel so frustrated.
System	Initiative	Action	The system requests for information related to the problem or provides suggestions and information for helping the user solve the problem.	How are your feelings at that time? Deep breaths can help people calm down. Some researches has found that ...
User	Non-Initiative	Feedback	The user responds to the system's request or delivers opinions on the system's statement.	Okay, this makes me feel better. No, I haven't.
System	Non-Initiative	Reflection	The system conveys the empathy to the user's emotion or shares similar experiences and feelings to comfort the user.	I understand you. I would also have been really frustrated if that happened to me. I'm sorry to hear about that.

Table 1: Definition and Examples for EAFR Schema Reflecting Patterns of Initiative Switch between Dialogue Participants in Emotional Support Conversations.

$$\text{Rel} = \frac{1}{\sum_{i=1}^{n} \mathcal{I}(r_i = S)} \sum_{i=1}^{n} \text{Rel}_i[r_i = S] \quad (5)$$

represents the change of the user's emotion intensity. $e_{<i}[r_{<i} = U]$ and $e_{>i}[r_{>i} = U]$ denote the emotion intensity of the first user utterance before and after the utterance i, respectively.

3.2 Analysis of Mixed Initiative in ESC

To reveal the necessity of incorporating mixed initiative into ESC systems, we analyze the different interaction patterns between empathetic dialogues (ED) and emotional support conversations (ESC): (i) EMPATHETICDIALOGUES (Rashkin et al., 2019), a dataset for ED that aims to provide empathetic responses for comforting the help-seeker, and (ii) ESConv (Liu et al., 2021), a dataset for ESC that aims to not only reduce users' emotional distress, but also help them understand and overcome the issues they face.

Due to the space limitation, we present the detailed analysis in Appendix A, including (i) the visualization of dialogue flow that indicates the initiative patterns between the user and system (A.2); (ii) the visualization of conversation progress that shows the phased change of the user's emotion intensity (A.3); and (iii) the evaluation of emotional support metrics that quantifies different aspects of mixed-initiative interactions (A.4).

3.3 Challenges of Mixed Initiative in ESC

The preliminary analysis reveals the importance of mixed-initiative interactions in ESC systems. Meanwhile, it is also challenging to balance the mixed-initiative interactions, as overacting in one way or taking the initiative inappropriately can be harmful to the emotional support conversations. Based on these analyses, we identify three key challenges in building a mixed-initiative ESC system:

***1) When* should the system take the initiative during the conversation?** The analysis of conversation progress (A.3) shows that taking initiative at different phases of the conversation may lead to different impacts on the user's emotional state. In particular, support strategies or dialogue acts attach great importance to conversational effectiveness in ESC (Zhang and Danescu-Niculescu-Mizil, 2020; Tu et al., 2022). Therefore, it is a crucial capability for the ESC system to determine whether to take the initiative at each conversation turn.

***2) What* kind of information is required for the system to initiate a subdialogue?** The analysis of mixed initiative metrics (A.4) show that the initiative system utterances are much informative than the non-initiative ones. Therefore, it is of great importance to discover necessary information and knowledge to make an appropriate mixed-initiative interaction. Researchers (Burleson, 2003) in communication and sociology states that the helpfulness of supportive statement is contingent on the following knowledge: (i) *Affective* Knowledge, the emotion recognition of the user's affective state, (ii) *Causal* Knowledge, the emotional reasoning of stressors that cause the current affective state of the user, and (iii) *Cognitive* Knowledge, the cognitive analysis of coping processes to solve the core problematic situation that the user faces.

***3) How* could the system facilitate the mixed-initiative interactions?** Since the system in ESC ultimately provides a natural language utterance to interact with the user, this challenge can be defined as a function that generates an initiative-aware utterance based on the given information.

3.4 Problem Definition

Similar to the ED problem, the ESC problem is typically defined as: given the dialogue context

$\mathcal{C} = \{u_1, u_2, ..., u_t\}$ and the description of the user's problematic situation s, the goal is to estimate a function $p(r|\mathcal{C}, s)$ that generates the target response r. In the light of the challenges discussed in Section 3.3, we further define the mixed-initiative emotion support conversation problem with the following three sub-tasks, corresponding to the above three challenges:
1) *Strategy Prediction* predicts the support strategy y that can be regarded as the fine-grained initiative.
2) *Knowledge Selection* selects appropriate knowledge k from the available resources $\mathcal{K}$.
3) *Response Generation* generates the mixed-initiative response r based on the predicted strategy and the selected knowledge.

4 Method

Motivated by the analysis in the last section, we propose the KEMI framework that aims to generate mixed-initiative responses with external knowledge. As illustrated in Figure 2, KEMI contains two parts: 1) Knowledge Acquisition, and 2) Mixed-initiative Response Generation.

4.1 Knowledge Acquisition

Commonsense knowledge is widely adopted to enhance the emotion reasoning in ESC systems. Despite the wide usage of commonsense knowledge in ESC systems, it is usually succinct and lacks specific context information. We propose an approach to retrieve relevant actual cases of ESC from a large-scale mental health knowledge graph, namely `HEAL` (Welivita and Pu, 2022), for compensating the deficiency of commonsense knowledge.

4.1.1 Query Expansion with COMET

Given the user utterance u_t at the current turn t, a straight-forward knowledge acquisition approach is to use u_t as the query to directly retrieve actual cases from the `HEAL` KG. However, there is limited information provided by the user utterance, which may hinder the preciseness and explainability of the knowledge retrieval. To this end, we exploit `COMET` (Bosselut et al., 2019), a commonsense knowledge generator, to expand the query with multi-perspective additional information regarding the user's affective and cognitive state.

Specifically, the current user utterance u_t is fed into `COMET` with five special relation tokens, $p \in \{$`[xReact]`, `[xIntent]`, `[xWant]`, `[xNeed]`, `[xEffect]`$\}$, to generate commonsense inference c_p for the relation p, *i.e.*, $c_p = \texttt{COMET}(p, u_t)$. Definitions of each commonsense relation can be found in Appendix B. Then the original user utterance u_t can be expanded with commonsense knowledge $\{c_p\}$.

4.1.2 Query Graph Construction

The actual case in HEAL (Welivita and Pu, 2022) is represented as a graph structure. Specifically, we consider 4 out of 5 types of nodes in `HEAL` that are related to response generation: *1) expectation*: commonly asked questions by the user in an emotional support conversation; *2) affective state*: emotional states associated with each speaker; *3) stressor*: the cause of emotional issues; and *4) response*: frequent types of responses by the system to address the user's problems. Edges are constructed to build the connections between nodes according to actual emotional support conversations. More details of `HEAL` can be found in Appendix C.

In accordance with the `HEAL` knowledge graph, the relation `[xReact]`, which reveals the user's emotional state, provides the same information as nodes in `HEAL` with the type of *affective state*. The relation `[xIntent]`, which reveals the causes of the user's current situation, also shares the same information as nodes in `HEAL` with the type of *stressor*. The rest of relations, including `[xWant]`, `[xNeed]`, and `[xEffect]`, which reveal the user's cognitive state, are relevant to the *responses* for addressing the user's problem. Therefore, the expanded query $\hat{u}_t = \{u_t, \{c_p\}\}$ can be represented as a graph with abstractive entity descriptions, as shown in Figure 2.

4.1.3 Subgraph Retrieval

To avoid enumerating all the subgraphs in `HEAL`, which is a densely-connected graph (over 2 million subgraphs), we propose a subgraph retrieval approach to select the top relevant subgraphs to form a candidate set. We first retrieve top-K entities relevant to each abstractive entity description in the expanded query graph $\hat{u}_t$. Specifically, we use sentence-BERT (Reimers and Gurevych, 2019) to be an embedding-based retriever $f_r(\cdot)$ for modeling the semantic similarity between the entities in the query and `HEAL`. With the retrieved top-K entities for each type of nodes, we merge them based on the edge connections in the knowledge graph to induce candidate subgraphs. Finally, we adopt top-N candidate subgraphs as the retrieved knowledge $\mathcal{K}$. The subgraphs are ranked by the sum of similarity scores of each node in the subgraph

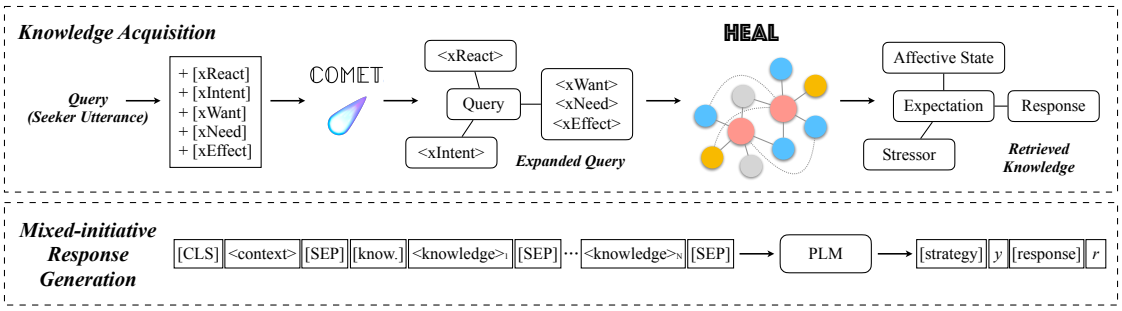


Figure 2: Overview of KEMI. Each expanded query is represented as a graph to retrieve subgraphs from HEAL, and each subgraph in HEAL can be regarded as an actual case of emotional support conversations.

$E = \{e_{\text{exp}}, e_{\text{aff}}, e_{\text{str}}, e_{\text{resp}}\}$:

$$\begin{aligned}\textbf{Sim}(\hat{u}_t, E) =& f_r(u_t, e_{\text{exp}}) + f_r(c_{\text{xR}}, e_{\text{aff}}) + f_r(c_{\text{xI}}, e_{\text{str}}) \\ &+ f_r([c_{\text{xW}}, c_{\text{xN}}, c_{\text{xE}}], e_{\text{resp}}).\end{aligned} \tag{6}$$

4.2 Mixed-initiative Response Generation

Given the dialogue context $\mathcal{C}$ and the retrieved knowledge $\mathcal{K}$, we first encode them into distributed representations with contextualized encoders. Specifically, we add special tokens to differentiate the roles of user and system as well as different types of knowledge as:

$$\texttt{<context>} = \texttt{[situ.]}, s, \texttt{[usr]}, u_1, \texttt{[sys]}, u_2, \ldots$$

$$\texttt{<know.>} = \texttt{[xR.]}, c_{\text{xR}}, \texttt{[xI.]}, \ldots, \texttt{[Aff.]}, e_{\text{aff}}, \ldots$$

Pretrained language models (PLMs), *e.g.*, GPT2 (Radford et al., 2019), have shown superior capability of generating high-quality responses in many dialogue systems, especially those PLMs pretrained on dialogue corpus, *e.g.*, BlenderBot (Roller et al., 2021). To leverage the advantages of these generative PLMs, we reformulate the mixed-initiative emotional support conversation problem as a Seq2Seq problem, which linearizes the input and output as a sequence of tokens as follows:

$$X = \texttt{[CLS]}, \texttt{<context>}, \texttt{[know.]}, \texttt{<know.>}_i, \ldots$$

$$Y = \texttt{[strategy]}, y, \texttt{[response]}, r$$

where X and Y are the linearized input and output sequences for Seq2Seq learning. Then the model is trained to maximize the negative log likelihood:

$$\mathcal{L} = -\frac{1}{L}\sum_{l=1}^{L} \log P(Y_l | Y_{<l}; X). \tag{7}$$

5 Experiment

5.1 Experimental Setups

Datasets We adopt the following two datasets for the evaluation: (i) ESConv (Liu et al., 2021), an emotional support conversation dataset, contains 1,300 dialogues with 38,365 utterances and 8 types of support strategies. We adopt the original train/dev/test split; and (ii) MI (Pérez-Rosas et al., 2016), a motivational interviewing dataset, contains 284 counseling sessions with 22,719 utterances and 10 types of behavior strategies. We randomly split the dataset for train/dev/test by 8:1:1[3].

Evaluation Metrics As for automatic evaluation, we adopt Macro F1 as the strategy prediction metric. Following previous studies (Liu et al., 2021; Tu et al., 2022), Perplexity (PPL), BLEU-n (B-n), and ROUGE-L (R-L) are included for the evaluation of response generation.

Baselines We provide extensive comparisons with both non-PLM and PLM-based methods, including three Transformer-based methods (Transformer (Vaswani et al., 2017), MoEL (Lin et al., 2019), and MIME (Majumder et al., 2020)) and four BlenderBot-based methods (BlenderBot (Roller et al., 2021), BlenderBot-Joint (Liu et al., 2021), GLHG (Peng et al., 2022)[4] , and MISC (Tu et al., 2022)[5]). Details about these baselines can be found in Appendix D.

Implementation Details KEMI is based on the BlenderBot model (Roller et al., 2021). Following previous BlenderBot-based models (Liu et al., 2021; Peng et al., 2022; Tu et al., 2022), we adopt the small version[6] of BlenderBot in experiments. The learning rate and the warmup step are set to

[3]Since there is no speaker label in the MI dataset, it is only adopted for response generation evaluation while the analysis of mixed initiative is not applicable.

[4]Since GLHG leverages the problem type as an additional label, we also report the ablation result for a fair comparison, *i.e.*, GLHG w/o $\mathcal{L}_2$ Loss.

[5]Due to a different train/test split adopted in Tu et al. (2022), we reproduce the performance of MISC on the standard split of ESConv (Liu et al., 2021).

[6]https://huggingface.co/facebook/blenderbot_small-90M

Model	F1↑	PPL↓	B-2↑	B-4↑	R-L↑
Transformer* (Vaswani et al., 2017)	-	81.55	5.66	1.31	14.68
MoEL* (Lin et al., 2019)	-	62.93	5.02	1.14	14.21
MIME* (Majumder et al., 2020)	-	43.27	4.82	1.03	14.83
BlenderBot** (Roller et al., 2021)	-	16.23	5.45	-	15.43
GLHG* (Peng et al., 2022)	-	**15.67**	7.57	2.13	16.37
GLHG w/o $\mathcal{L}_2$ Loss* (Peng et al., 2022)	-	-	6.15	1.75	15.87
BlenderBot-Joint (Liu et al., 2021)	19.23	16.15	5.52	1.29	15.51
MISC (Tu et al., 2022)	19.89	16.08	7.62	2.19	16.40
KEMI	**24.66**†	15.92	**8.31**†	**2.51**†	**17.05**†

Table 2: Experimental results on ESConv. * and ** indicate the results reported in Peng et al. (2022) and Liu et al. (2021) respectively. Other results are reproduced. † indicates statistically significant improvement (p<0.05) over the best baseline.

Model	F1↑	PPL↓	B-2↑	B-4↑	R-L↑
Transformer (Vaswani et al., 2017)	-	65.52	6.23	1.52	15.04
BlenderBot (Roller et al., 2021)	-	16.06	6.57	1.66	15.64
BlenderBot-Joint (Liu et al., 2021)	22.66	14.74	7.28	2.18	16.41
MISC (Tu et al., 2022)	22.68	14.33	7.75	2.30	17.11
KEMI	**25.91**†	**13.84**†	**8.52**†	**2.72**†	**18.00**†

Table 3: Experimental results on MI Counseling.

be 3e-5 and 100, respectively. The max input sequence length and the max target sequence length are 160 and 40, respectively. We retrieve the top-1 subgraph from HEAL as the knowledge. The training epoch is set to 5 and the best model is saved according to the PPL score in the dev set.[7]

5.2 Overall Performance

Table 2 and Table 3 summarize the experimental results on the ESConv and MI dataset, respectively. Among the baselines, BlenderBot-based methods largely outperform Transformer-based methods by leveraging the valuable pretrained knowledge. GLHG and MISC effectively exploit the commonsense knowledge to improve the performance of response generation. Besides, the joint learning with strategy prediction task is beneficial to the performance of response generation. Finally, KEMI substantially outperforms other methods with a noticeable margin. This indicates the domain-specific actual case knowledge from HEAL can alleviate the reliance on large-scale PLMs. Compared with commonsense knowledge, the knowledge from HEAL is much more effective in predicting support strategies, as this relevant knowledge can serve as an real example for guiding the system to respond.

[7] https://github.com/dengyang17/KEMI

vs.	BlenderBot-Joint			MISC		
	Win	Tie	Loss	Win	Tie	Loss
Flu.	26%	**51%**	23%	37%	**47%**	16%
Ide.	**50%**	38%	12%	**46%**	30%	24%
Com.	**46%**	40%	14%	**44%**	30%	26%
Sug.	**52%**	22%	26%	**52%**	16%	28%
Ove.	**62%**	20%	18%	**70%**	12%	18%

Table 4: Human evaluation results (KEMI vs.).

5.3 Human Evaluation

Following previous studies (Liu et al., 2021; Peng et al., 2022), we conduct human evaluation to compare the generated responses from two given models on five aspects: 1) *Fluency*: which model's response is more fluent? 2) *Identification*: which model's response is more skillful in identifying the user's problem? 3) *Comforting*: which model's response is better at comforting the user? 4) *Suggestion*: which model can give more helpful and informative suggestions? 5) *Overall*: which model's response is generally better? We randomly sample 100 dialogues from ESConv and three annotators are asked to determine the *Win/Tie/Lose* for each comparison.

Table 4 presents the human evaluation results. We compare the generated responses from KEMI with those produced by other two baselines, BlenderBot-Joint and MISC. The results show that KEMI achieves remarkable improvement on initiative interactions, including *Identification* and *Suggestion*. Consequently, KEMI can generate more satisfactory and helpful responses than other methods, according to the *Overall* metric.

5.4 Ablation Study

In order to investigate the effect of each sub-task and each type of knowledge on the final performance, we report the experimental results of the ablation study in Table 5. In general, both the strategy prediction and the knowledge selection tasks as well as all types of knowledge contribute to the final performance more or less. There are several notable observations in detailed comparisons: (i) The knowledge from HEAL is the key to the improvement on the strategy prediction task, since the actual case knowledge can provide a good guidance for the next support strategy. (ii) Different from discarding the actual case knowledge (w/o HEAL), discarding the commonsense knowledge

Strategy	Knowledge	F1↑	PPL↓	B-2↑	R-L↑
-	-	-	16.23	5.45	15.43
-	KEMI	-	16.16	6.54	16.21
Joint	KEMI	24.66	15.92	8.31	17.05
Joint	w/o COMET	23.26	15.74	7.60	16.47
Joint	w/o HEAL	19.99	16.08	7.98	16.92
Joint	w/o *Affective*	22.68	16.08	8.22	16.98
Joint	w/o *Causal*	23.14	15.94	8.16	16.92
Joint	w/o *Cognitive*	20.24	16.22	7.62	16.64
Joint	Oracle	**32.38**	12.79	18.45	28.01
Oracle	KEMI	-	15.92	9.75	18.81
Oracle	Oracle	-	**12.78**	**19.11**	**28.88**

Table 5: Ablation study. Oracle knowledge is obtained by the lexical match between the reference response and the candidate knowledge from HEAL.

	Proactivity		Information			Repetition			Relaxation		
	Init.	Non.	Init.	Non.	All	Init.	Non.	All	Init.	Non.	All
BB	0.36	**0.64**	1.79	1.32	1.48	1.00	1.11	1.07	-0.01	0.11	0.07
BB-J	**0.68**	0.32	1.89	1.18	1.66	1.18	1.09	**1.15**	0.01	0.07	0.03
MISC	0.61	0.39	1.91	1.25	1.65	1.16	**1.12**	1.14	0.00	0.04	0.02
KEMI	0.45	0.55	**2.04**	**1.40**	**1.68**	**1.18**	1.09	1.13	**0.09**	**0.13**	**0.11**
REF	0.51	0.49	3.09	3.01	3.05	1.12	1.06	1.09	0.10	0.13	0.11

Table 6: Emotional support metrics. BB and BB-J denote BlenderBot and BlenderBot-Joint.

(w/o COMET) brings a positive effect on the fluency metrics (PPL), as the commonsense knowledge is not a natural sentence. However, the COMET contributes more on the content-preserving metrics (BLEU and ROUGE) than the HEAL, indicating that the succinct commonsense knowledge can be more precise. (iii) Among the three types of knowledge, cognitive knowledge is the most effective one for both strategy prediction and response generation tasks. (iv) Using Oracle strategy and Oracle knowledge substantially improves the overall performance, which demonstrates the effectiveness of considering these two sub-tasks in ESC systems. The performance gap between KEMI and Oracle also shows that the knowledge selection is very challenging and there is still much room for improvement.

5.5 Analysis of Mixed Initiative

We conduct the mixed initiative analysis introduced in Section 3.2 over the proposed KEMI method and other baselines. Since the calculation of the Relaxation metric in Eq.(4) requires the emotion intensity score of the user feedback, we adopt a model-based user simulator for automatic evaluation, which is described in Appendix A.1.3.

5.5.1 Emotional Support Metrics

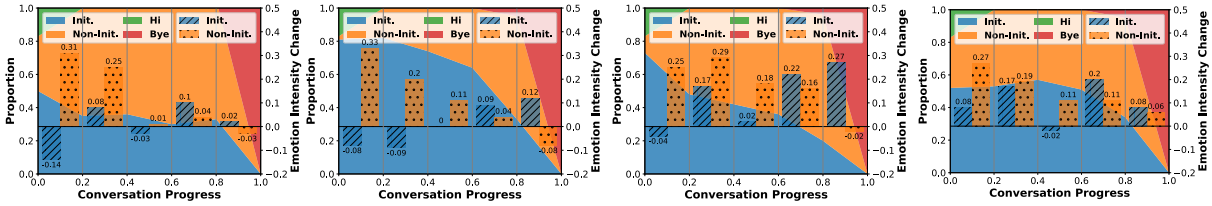
Table 6 summarizes the results of the four emotional support metrics for the generated responses from four BlenderBot-based methods and the reference responses in the test set. Note that, for a fair comparison, we also adopt Eq.(9) to calculate the Relaxation metric for the reference responses in the test set (*i.e.*, REF). It can be observed that (i) As for the Proactivity metric, BlenderBot tends to act passively in ESC. While BlenderBot-Joint and MISC overly take the initiative after simply taking into account the support strategies. KEMI effectively balances the proportion of initiative and non-initiative interactions in ESC. (ii) With the actual case knowledge, KEMI can generate much informative responses than other baselines w.r.t the Information metric. However, there is still a large gap to reach the reference responses. (iii) Indeed, it is relatively easier to generate responses that repeat the previous information w.r.t the Repetition metric. (iv) KEMI outperforms other baselines in terms of the Relaxation metric on the initiative interactions with a large margin, which shows the superiority of KEMI on taking the initiative role for helping the user to solve emotional problems.

5.5.2 Conversation Progress

We conduct the conversation progress analysis by dividing the whole conversation into five equal length intervals and observing the change of users' emotion intensity levels at each conversation phase. As the results shown in Figure 3, we observe that BlenderBot and MISC have a clear inclination to take non-initiative and initiative interactions in all stages of the conversation, respectively. Our KEMI method shares a more similar progress as the reference conversation with a balanced interaction pattern. More importantly, the initiative responses generated by KEMI has a more positive impact on the user's emotional intensity than other baselines, especially in the last two stages of the conversation. This result indicates that KEMI effectively takes the initiative to generate responses that can provide suggestions or information for relaxing the help-seekers by solving their emotional problems.

5.6 Case Study

To intuitively show the superiority of KEMI over other baselines, Figure 4 presents a case study of generated responses with the scores of mixed-initiative metrics. In the reference response, the system takes the initiative to provide useful sug-



(a) BlenderBot (b) MISC (c) KEMI (Ours) (d) ESConv (Reference)

Figure 3: The distribution of system utterance initiative (the stack plot) and the user's emotion intensity change (the bar chart) at different conversation progress. Higher scores of the emotion intensity change represent better emotion improvement of users.

Dialogue Context	Model	Response (Knowledge)	Pro	Inf	Rep	Rel
Situation: My set of friends with include 6 close women are falling apart because of not communicating often and not being physically close to one another. **Supporter**: Hello, what seems to be on your mind? **Seeker**: Friends and our relationships amongst one another. Two are arguing and I want to help both. We only can facetime each other because of Corona and it is taking a toll among our group. … **Seeker**: How do I make her understand this? **Supporter**: I think it is great that you want to actively help, it says a lot about you. **Seeker**: Thanks! What should I say to her?	REF	Sometimes the best **way** to *help* someone *understand* is to *give* them a little **time** and **space**. **Trying** to **force** someone to *understand* doesn't always **work**.	I	6	4	0.51
	BB	I would *say* that you are **trying** to *help* her and you are **trying** to *help* her.	N	1	2	0.20
	BB-J	I would *tell* her that you are sorry and that you want to *help* her.	N	0	2	0.21
	MISC	I *think* you should be able to *tell* her that you are **trying** to *help* her and that you are **trying** to *help* her. ([xReact] sad [xIntent] to be heard [xWant] to get her to understand [xNeed] none [xEffect] none)	I	1	3	0.27
	KEMI	I *think* you should *give* her **time** and **space**. It is important to *give* her **time** and **space** to *understand* your **feelings**. ([aff] Sad [str] parents, tell, telling, bring, open [resp] You just need to give her time to understand that.)	I	3	3	0.34

Figure 4: Case study. **Bold** terms denote new (red) and *repeated* (blue) frequent terms respectively.

gestions to the user for solving her/his problem, which effectively reduce the user's emotional intensity. Among the generated responses, BlenderBot and BlenderBot-Joint decide to convey empathy to the user by paraphrasing the previous information, while MISC and KEMI proactively initiate a discussion about potential solutions to the problem. Based on the Relaxation metric, two initiative responses can better comfort the emotional intensity of the user than two non-initiative responses. Furthermore, KEMI can generate more informative and specific responses with actual case knowledge.

6 Conclusions

In this paper, we design a novel analysis framework for analyzing the feature of mixed initiative in ESC. The analysis demonstrates the necessity and importance of mixed-initiative interactions in ESC systems. To this end, we propose the KEMI framework to tackle the problem of mixed-initiative ESC. KEMI first retrieves actual case knowledge from a large-scale mental health knowledge graph with query expansion and subgraph retrieval. Then KEMI performs multi-task learning of strategy prediction and response generation with the retrieved knowledge. Extensive experiments show that KEMI outperforms existing methods on both automatic and human evaluation. The analysis also shows the effectiveness of incorporating actual case knowledge and the superiority of KEMI on the mixed-initiative interactions.

Limitations

In this section, we analyze the limitations of this work:

- As it is the first attempt to analyze the mixed-initiative interactions in emotional support conversations, the proposed metrics can be further improved for more robust evaluation.
- Since the knowledge retrieval is not the focus of this work, we did not spend much space on discussing the choice of different retrieval methods. As shown in Table 5, there is still much room for improving the knowledge retrieval from a large scale knowledge graph. It is also worth studying more efficient retrieval methods for retrieving knowledge from a densely connected KG.
- The proposed method requires an additional mental health related knowledge graph constructed by experts or knowledgeable workers, which is probably difficult to obtain in some applications. However, different from other knowledge-intensive tasks that can be benefited from open-domain knowledge (*e.g.,* Wikipedia), it attaches

great importance in the professionals of the knowledge for building a helpful and safe ESC system.

Ethical Considerations

The datasets adopted are publicly available and widely studied benchmarks collected from professionals or well-trained annotators. All personally identifiable and sensitive information, *e.g.*, user and platform identifiers, in these dataset has been filtered out. We do not make any treatment recommendations or diagnostic claims. Compared with existing methods for emotional support conversations, the proposed method can be regarded as one step further to a more safer ESC system. The proposed method retrieves knowledge from a well-established mental health knowledge graph, which can be maintained by filtering out harmful information when applying into applications. Then the knowledge-enhanced approach can alleviate the randomness during the response generation and provide the guidance towards more positive responses. In order to prevent the happening of unsafe cases, the analysis of emotion intensity prediction can also serve as an alarming mechanism that calls for handoffs to an actual psychologist.

References

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *SIGIR 2019*, pages 475–484.

James E Allen, Curry I Guinn, and Eric Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Trans. Assoc. Comput. Linguistics*, 4:463–476.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL 2019*, pages 4762–4779.

Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge.

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. *CoRR*, abs/2210.04242.

Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. 2023. A survey on proactive dialogue systems: Problems, methods, and prospects. *CoRR*, abs/2305.02750.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022a. PACIFIC: towards proactive conversational question answering over tabular and textual data in finance. In *EMNLP 2022*, pages 6970–6984.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *SIGIR 2021*, pages 1431–1441.

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022b. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *WWW '22: The ACM Web Conference 2022*, pages 2998–3008.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of ACL: EMNLP 2021*, pages 807–819.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *EMNLP 2021*, pages 2227–2240.

Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of proactive dialogue strategies on human-computer trust. In *UMAP 2020*, pages 107–116.

Wenqiang Lei, Yao Zhang, Feifan Song, Hongru Liang, Jiaxin Mao, Jiancheng Lv, Zhenglu Yang, and Tat-Seng Chua. 2022. Interacting with non-cooperative user: A new paradigm for proactive dialogue policy. In *SIGIR 2022*.

Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2022a. Diaasq : A benchmark of conversational aspect-based sentiment quadruple analysis. *CoRR*, abs/2211.05705.

Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. 2022b. Knowledge bridging for empathetic dialogue generation. In *AAAI 2022*.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *SIGIR 2021*, pages 2041–2045.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP 2019*, pages 121–132.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *ACL/IJCNLP 2021*, pages 3469–3483.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Usha Lokala, Aseem Srivastava, Triyasha Ghosh Dastidar, Tanmoy Chakraborty, Md Shad Akthar, Maryam Panahiazar, and Amit P. Sheth. 2022. A computational approach to understand mental health from reddit: Knowledge-aware multitask learning framework. In *ICWSM 2022*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: mimicking emotions for empathetic response generation. In *EMNLP 2020*.

Ganeshan Malhotra, Abdul Waheed, Aseem Srivastava, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In *WSDM 2022*.

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. *CoRR*, abs/2204.12749.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence C. An. 2016. Building a motivational interviewing dataset. In *CLPsych@NAACL-HLT 2016*, pages 42–51.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence C. An, Kathy J. Goggin, and Delwyn Catley. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *EACL 2017*, pages 1128–1137.

Betty Pfefferbaum and Carol S North. 2020. Mental health and the covid-19 pandemic. *New England Journal of Medicine*, 383(6):510–512.

Konigari Rachna, Saurabh Ramola, Vijay Vardhan Alluri, and Manish Shrivastava. 2021. Topic shift detection for mixed initiative response. In *SIGdial 2021*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL 2019*, pages 5370–5381.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP-IJCNLP 2019*, pages 3980–3990.

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 *EACL 2021*, pages 300–325.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: commonsense-aware empathetic response generation. *CoRR*, abs/2109.05739.

Tulika Saha, Saraansh Chopra, Sriparna Saha, Pushpak Bhattacharyya, and Pankaj Kumar. 2021. A large-scale dataset for motivational dialogue system: An application of natural language generation to mental health. In *IJCNN 2021*, pages 1–8.

Tulika Saha, Saichethan Miriyala Reddy, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Mental health disorder identification from motivational conversations. *IEEE Transactions on Computational Social Systems*.

Ivan Sekulic, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *WSDM 2022*, pages 888–896.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW 2021*, pages 194–205.

Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *ACL 2022*, pages 3096–3107.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *SIGdial 2020*, pages 10–20.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *ACL 2019*, pages 3685–3695.

Jina Suh, Eric Horvitz, Ryen W. White, and Tim Althoff. 2021. Population-scale study of human needs during the COVID-19 pandemic: Analysis and implications. In *WSDM 2021*, pages 4–12.

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *ACL 2022*, pages 2952–2973.

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 *ACL 2022*, pages 308–319.

Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2020. An analysis of mixed initiative and collaboration in information-seeking dialogues. In *SIGIR 2020*, pages 2085–2088.

Svitlana Vakulenko, Evangelos Kanoulas, and Maarten de Rijke. 2021. A large-scale analysis of mixed initiative in information-seeking dialogues for conversational search. *ACM Trans. Inf. Syst.*, 39(4):49:1–49:32.

Svitlana Vakulenko, Kate Revoredo, Claudio Di Ciccio, and Maarten de Rijke. 2019. QRFA: A data-driven model of information-seeking dialogues. In *ECIR 2019*, pages 541–557.

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 *NeurIPS 2017*, pages 5998–6008.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *CIKM 2019*, pages 1401–1410.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *COLING 2020*, pages 4886–4899.

Anuradha Welivita and Pearl Pu. 2022. Heal: A knowledge graph for distress management conversations. In *AAAI 2022*.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *EMNLP 2021*, pages 1251–1264.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *ACL 2019*.

Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *WWW 2020*, pages 418–428.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *ACL 2020*, pages 5276–5289.

Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *KDD 2020*, pages 1512–1520.

Wenxuan Zhang, Yang Deng, Xin Li, Lidong Bing, and Wai Lam. 2021a. Aspect-based sentiment analysis in question answering forums. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4582–4591.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021b. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 9209–9219.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021c. Towards generative aspect-based sentiment analysis. 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*, pages 504–510.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *CoRR*, abs/2203.01054.

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 2021*, pages 14577–14585.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI 2018*, pages 730–739.

Appendix

A Analysis of Mixed Initiative

A.1 Tools for Analysis

We introduce two models that are adopted as off-the-shelf tools for the analysis of mixed initiative.

A.1.1 Utterance Initiative Classification

To facilitate automatic analysis of utterance initiative, we train two utterance classification models by fine-tuning the pre-trained RoBERTa$_{\text{large}}$ model (Liu et al., 2019) on the ESConv (Liu et al., 2021) dataset, one for system utterance classification and the other for user utterance classification. We concatenate the previous utterance from another participant and the current utterance as the input for the binary initiative classification, either Initiative (I) or Non-Initiative (N). However, there is no initiative label in ESConv. Therefore, we manually annotate the initiative labels, I or N, for each utterance according to the EAFR schema. The resulting dataset contains ∼38K utterance-label pairs (E: 13K, A: 9K, F: 7K, R: 10K).

A.1.2 Emotion Intensity Prediction

Similarly, we also fine-tune an emotion intensity prediction model $f_e(\cdot)$ based on the pre-trained RoBERTa$_{\text{large}}$ model (Liu et al., 2019) on the ESConv (Liu et al., 2021) dataset. Given a user utterance, the model aims to predict the negative emotion intensity level $e_{ij} = f_e(u_{ij})$, ranging from 1 to 5, which indicates the user's emotional state. In ESConv, the initial and final emotion intensity levels of the user have already been annotated. Therefore, we regard the first user utterance after greetings to match with the initial emotion intensity, while the last user utterance before goodbyes to match with the final emotion intensity. The resulting dataset contains 2,450 utterance-label pairs (1: 331, 2: 506, 3: 557, 4: 629, 5: 427).

A.1.3 User Simulator

Inspired by the evaluation of mixed-initiative CIS (Sekulic et al., 2022), we simulate a user based on a large-scale generative language model, namely BlenderBot (Roller et al., 2021). In our case, we fine-tune a semantically-conditioned generation model $g(\cdot)$, guided by the underlying problematic situation:

$$p_g(\boldsymbol{a}|s, \mathcal{C}, r) = \prod_{l=1}^{L} p_g(a_l|a_{<l}; s, \mathcal{C}, r), \quad (8)$$

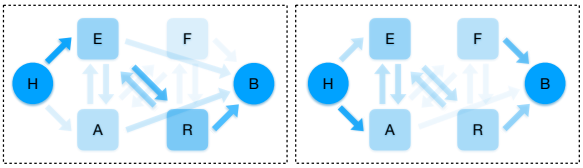
where a is the user's feedback to the generated response r. The generation model is fine-tuned on the whole dataset, including the test set. If the ESC system generates a perfect response, the user simulator should give the ground-truth feedback as the real user.

We adopt the same utterance initiative classification model and emotion intensity prediction model $f_e(\cdot)$ described in Appendix A.1.2 to annotate the generated response. The annotation results are used for calculating the emotional support metrics. In particular, the calculation of Relaxation metric involves the user's emotion intensity after receiving the generated response. The user simulator $g(\cdot)$ is employed to simulate the user's feedback. Then the calculation of the Relaxation metric in Eq.(4) becomes:

$$\text{Rel}_i[r_i = S] = f_e(u_{<i}[r_{<i} = U]) - f_e(g(s, \mathcal{C}, r)). \quad (9)$$

A.2 Dialogue Flow

Following previous studies on mixed-initiative CIS systems (Vakulenko et al., 2019, 2021), we draw the dialogue flow diagram to observe the initiative switch patterns between dialogue participants in ED and ESC. As shown in Figure 5, the circles represent the beginning and ending of the dialogue, while the boxes represent the EAFR utterance labels. The color intensity denotes the proportions of the utterance labels and the initiative transitions. There are several notable observations: (i) As for the proportion of EAFR labels, *Expression* and *Reflection* constitute the majority of the utterances in ED, while four labels are more equally distributed in ESC. (ii) As for the beginning and ending of dialogues, users are more often to take the initiative to start a conversation in ED, and the dialogue will be ended by the system. Differently, in ESC, the conversation is usually started by the system. (iii) As for the initiative switches, most of cases in ED are that users express their feelings and then the system tries to comfort them with empathy. However, the proportion of each type of initiative transitions in ESC is relatively equal. Therefore, we conclude that **the system in ED generally serves as a passive role, while the system in ESC needs to switch the initiative role during the conversation**.

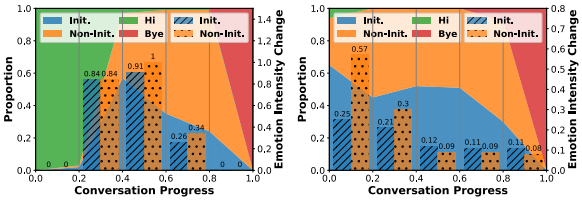


(a) Empathetic Open-domain Conversations (b) Emotional Support Conversations

Figure 5: Dialogue flow. H, B, E, A, F, and R denote Hi, Bye, Expression, Action, Feedback, and Reflection. The color intensity denotes the proportions of the utterance labels and the initiative transitions.

A.3 Conversation Progress

We analyze the conversation progress by dividing the whole conversation into five equal length intervals. To alleviate the noise from greeting (Hi) and farewell (Bye) utterances, we heuristically identify these utterances by rules, *e.g.*, containing "Hi/Hello" at the beginning or "Bye/Goodbye" at the end of the conversation. Specifically, we compute the distribution of initiative labels for system utterances and the average change of emotion intensity levels at each conversation phase. As shown in Figure 6, under both cases, the system tends to take the initiative at the beginning of the conversation for exploring the user's problem, while acting passively at the latter stage of the conversation.



(a) EmpatheticDialogues (b) ESConv

Figure 6: The distribution of utterance initiative (the stack plot) and the emotion intensity change (the bar chart) at different conversation progress.

	Proactivity		Information			Repetition			Relaxation		
	Init.	Non.	Init.	Non.	All	Init.	Non.	All	Init.	Non.	All
ED	0.28	0.72	2.14	2.69	2.46	0.42	0.44	0.43	0.83	0.82	0.83
ESC	0.48	0.52	3.32	3.06	3.19	1.06	1.18	1.12	0.16	0.20	0.18

Table 7: Comparisons on emotional support metrics.

Interestingly, at the early phase of conversations, compared with non-initiative utterances, system-initiative ones fail to relax the emotion intensity of help-seekers in ESC. This is because the request for information from users to understand their problems is likely to raise users' negative emotions. However, at the latter stage of the conversation, initiative utterances can better lower down users' intensity levels, leading to a higher emotional intensity change rate than non-initiative ones. This indicates that **(i) the timing for system-initiative interactions is important, and (ii) it is more helpful to provide suggestions or information for users to solve the problem when the emotion of users has been eased**.

A.4 Emotional Support Metrics

Table 7 summarizes the scores of the emotional support metrics introduced in Section 3.1 for two datasets. Firstly, the proportion of system-initiative interactions in ESC is much higher than that in ED, showing the importance of mixed initiative in ESC systems. Secondly, system-initiative utterances provide more information than non-initiative utterances in ESC, while an opposite result is observed in ED. This shows that the ESC system provides informative responses when taking the initiative. Thirdly, in both datasets, there is more repetitive information in non-initiative system utterances than initiative ones, indicating that reflection is more important in non-initiative interactions. Last but not least, the average *relaxation* score in ESC is much lower than that in ED. We attribute this to two reasons: (i) Empathetic responses have more positive effects on the user's emotions. (ii) The system-initiative interactions sometimes may increase the user's emotion intensity, as discussed in Appendix A.3.

	Stressor	Expectation	Response	Affect. State
Stressor	**4,363**	9,801	-	-
Expectation	9,801	**3,050**	26,628	3,050
Response	-	26,628	**13,416**	-
Affect. State	-	3,050	-	**41**

Table 8: Statistics of HEAL adopted in our experiments.

B Definition of COMET Relations

We adopt five types of commonsense relations in COMET (Bosselut et al., 2019), whose original definitions are as follows:

- `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.

C Details of HEAL

HEAL (Welivita and Pu, 2022) is a knowledge graph developed upon 1M distress discussions and their corresponding consoling responses curated from mental health support conversations. It consists of 22K nodes with five different types: *stressors*, *expectations*, *responses*, *feedback*, and *affective states* associated with distress dialogues, and forms 104K connections between different types of nodes. The statistics of the adopted HEAL are presented in Table 8.

D Baselines

We provide extensive comparisons with the following strong baselines, including both non-PLM and PLM-based methods:

- Transformer (Vaswani et al., 2017) for Seq2Seq response generation.

- MoEL (Lin et al., 2019) is a Transformer-based model that involves multi-decoders to enhance the empathy for different emotions.
- MIME (Majumder et al., 2020) is a Transformer-based model that mimics the emotion of the speaker for empathetic response generation.
- BlenderBot (Roller et al., 2021) is an open-domain dialogue model pretrained with multiple skills, including empathetic responding. BlenderBot-Joint (Liu et al., 2021) jointly predicts strategies and generates responses.
- GLHG (Peng et al., 2022) is a BlenderBot-based model, which employs a hierarchical graph network to encode multi-source information.
- MISC (Tu et al., 2022) is a BlenderBot-based model, which incorporates commonsense knowledge and mixed support strategy to jointly predicts support strategies and generates responses.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitation Section

☑ A2. Did you discuss any potential risks of your work?
Ethical Considerations Section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

Section 5

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5.1

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.