Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation

Wei Peng[1,2] , **Yue Hu**[1,2] * , **Luxi Xing**[1,2] , **Yuqiang Xie**[1,2] , **Yajing Sun**[1,2] , **Yunpeng Li**[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
{pengwei,huyue}@iie.ac.cn

Abstract

Emotional support conversation aims at reducing the emotional distress of the help-seeker, which is a new and challenging task. It requires the system to explore the cause of help-seeker's emotional distress and understand their psychological intention to provide supportive responses. However, existing methods mainly focus on the sequential contextual information, ignoring the hierarchical relationships with the global cause and local psychological intention behind conversations, thus leads to a weak ability of emotional support. In this paper, we propose a Global-to-Local Hierarchical Graph Network to capture the multi-source information (global cause, local intentions and dialog history) and model hierarchical relationships between them, which consists of a multi-source encoder, a hierarchical graph reasoner, and a global-guide decoder. Furthermore, a novel training objective is designed to monitor semantic information of the global cause. Experimental results on the emotional support conversation dataset, ESConv, confirm that the proposed GLHG has achieved the state-of-the-art performance on the automatic and human evaluations.

1 Introduction

Emotional Support Conversation (ESConv) task, which focuses on effectively providing support to a help-seeker [Burleson *et al.*, 2006; Heaney and Israel, 2008; Slovák *et al.*, 2015], is a new and challenge task [Liu *et al.*, 2021]. Different from Emotional Conversation task, where the dialog agent generates emotional responses with the given emotion [Zhou *et al.*, 2018], ESConv task expects the dialog system to possess a more high-level and complex ability with the target of reducing the emotional distress of the help-seeker and providing the supportive responses. It is significant to construct such a emotional support dialog system [Zhou *et al.*, 2020; van der Zwaan *et al.*, 2012] especially in social interactions (accompanying and cheering up the user), mental health support (comforting a frustrated help-seeker and helping identify
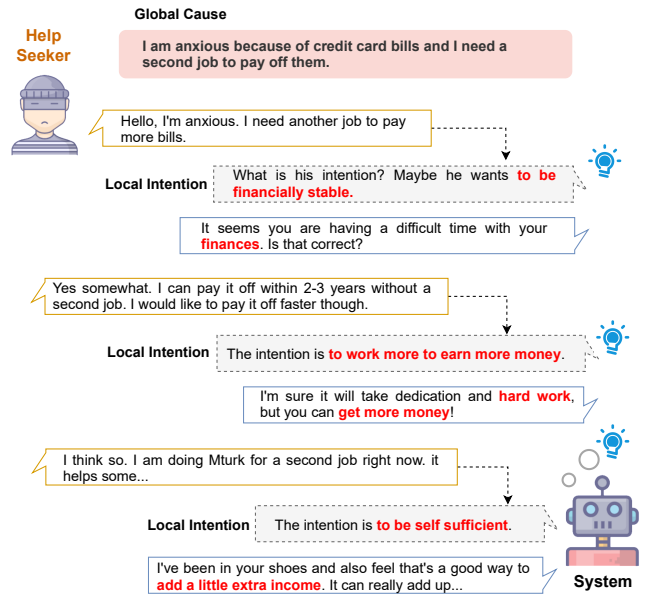
Figure 1: An example to show the emotional support conversation which consists of a help-seeker and system that provides the supportive responses. The red background is the global cause to describe the emotional problem and the gray background is the local intention to imply the current speaker's state. These factors are not explicitly mentioned by the help-seeker. Words in red reflect relevance.

the problem), customer service chats (appeasing an angry customer and providing solutions), etc [Liu *et al.*, 2021].

During the emotional support conversation, the system is required to explore the cause of the help-seeker's emotional problem and understand their psychological intentions to provide more supportive responses [Rains *et al.*, 2020]. On the one hand, the cause describes the reason for the stressful problem of the help-seeker [Liu *et al.*, 2021], which can globally control the whole flow of the emotional support conversation. On the other hand, psychological intention is implicitly expressed by the help-seeker, which can help the system locally understand the help-seeker's mental state of the current turn. However, the above factors will not be explicitly mentioned by the help-seeker. And existing methods [Liu *et al.*, 2021] mainly focus on the sequential contextual information, ignoring the hierarchical relationships between the global cause,

local psychological intention and dialog history, which leads to a weak ability of emotional support. An example in Figure 1 is illustrated to explain the above process. Globally, the system needs to capture the global cause (*The help-seeker is anxious because of credit card bills and needs a second job*) to enhance the semantic information of the whole conversation. Locally, the system should respond with the captured intention to understand the help-seeker's mental state during different turn. For example, in turn two, the help-seeker indicates that he/she wants to **work more to earn more money**. And the system generates an appropriate and encouraging response *It will take dedication and **hard work**, but you can **get more money***. Hence, 1) how to capture the global cause and local psychological intentions and 2) how to model the relationships between them are two important challenges in the emotional support conversation.

To address these problems, we propose a novel **G**lobal-to-**L**ocal **H**ierarchical **G**raph Network (**GLHG**) for emotional support modeling. For the first issue, a multi-source encoder is presented to capture the global cause and local psychological intention (with COMET, a pre-trained generative commonsense reasoning model [Bosselut *et al.*, 2019]). As for the second issue, a hierarchical graph reasoner is designed to model the hierarchical relationships between the global cause (dialog-level), local psychological intention (sentence-level) and dialog history. In addition, a novel training objective is designed in global-guide decoder to monitor semantic information of the global cause, so that the global cause will not be biased by the local intention.

The contributions can be summarized as follows:

- We propose a Global-to-Local Hierarchical Graph Network (GLHG) for emotional support conversation from the global-to-local perspective.
- To capture the global cause and local intention, the multi-source encoder utilizes the information of situation and incorporates psychological intention with COMET.
- To model the different level relationships, the hierarchical graph reasoner makes an interaction between the global cause, local psychological intention and dialog history.
- Experiments on the dataset show that the GLHG achieves the state-of-the-art performance in terms of both automatic evaluation metrics and human evaluations.

2 Related Work

2.1 Emotional & Empathetic Conversation

In recent years, research on emotional conversation has embraced a booming in dialog systems [Zhou *et al.*, 2018; Huang *et al.*, 2020; Peng *et al.*, 2022]. [Zhou *et al.*, 2018] propose Emotional Chatting Machine to accurately generate emotional responses conditioning on a manually specified label. Different from the task of emotional chatting conversation, the task of empathetic dialog generation [Lin *et al.*, 2019; Rashkin *et al.*, 2019] aim to respond to a speaker accordingly on the emotional situation in an empathetic way. Unlike the above tasks, emotional support conversation [Liu *et al.*, 2021] focuses on exploring the help-seeker's problems and generating more supportive responses.

2.2 Graph Modeling in Dialog

Motivated by impressive advancements and structured modeling of GCN [Guo *et al.*, 2021; Sun *et al.*, 2021], efforts towards better performance and construction utilizing GCNs have sprung up on dialog systems [Ghosal *et al.*, 2019; Qin *et al.*, 2021; Xu *et al.*, 2021]. DialogueGCN [Ghosal *et al.*, 2019] leverages self and inter-speaker dependency with GCN to model conversational context. [Qin *et al.*, 2021] propose a Co-Interactive Graph Attention Network for dialog act recognition and sentiment classification task. DVAE-GNN [Xu *et al.*, 2021] focuses on discovering dialog structure in the open-domain dialogue, and EGAE [Qin *et al.*, 2021] captures the dialogue schema in the task-oriented dialogue with the graph network. Unlike the previous work, we propose a Hierarchical Graph Reasoner for modeling the relationships between the global cause, local psychological intentions and dialog history, and to facilitate information flow over the graph.

2.3 Commonsense Knowledge

ATOMIC [Sap *et al.*, 2019] is a commonsense knowledge [Yu *et al.*, 2020; Xing *et al.*, 2020] graph, which focuses on inferential knowledge organized as typed if-then relations with variables (e.g., if "PersonX drinks coffee", then "PersonX needed to brew coffee"). The type of relations contains If-Event-Then-Mental-State, If-Event-Then-Event and If-Event-Then-Persona. In this paper, to capture the mental state of the help-seeker, we focus on the If-Event-Then-Mental-State which infers three commonsense relations: PersonX's reaction to the event (*xReact*), PersonX's intention before the event (*xIntent*) and PersonY's reaction to the event (*oReact*). Since oReact refers to the other person (e.g., system), which is not considered, we ignore it in this paper. To obtain the psychological intention (*xIntent*) of the help-seeker, we employ COMET [Bosselut *et al.*, 2019], a pre-trained generative commonsense reasoning model, to generate rich commonsense descriptions in natural language (e.g., input "I need another job to pay more bills", then output "to be financially stable").

3 Approach

As shown in Figure 2, the proposed GLHG consists of three parts. Firstly, **Multi-source Encoder** captures the global cause information and local psychological intentions, as well as encoding the dialog history. Secondly, **Hierarchical Graph Reasoner** incorporates the different level representations for graph reasoning from the global-to-local perspective. Finally, **Global-guide Decoder** generates the supportive responses and monitors semantic information of the global cause by the novel designed training objective.

3.1 Problem Formulation

The problem formulation of our task could be formulated in the following. Given a dialog history $C = (u^1, u^2, \ldots, u^{N-1})$, a set of $N - 1$ utterances, where $u_m^n = (x_1^n, x_2^n, \ldots, x_M^n)$ that consists of M words, with $s = (s_1, s_2, \ldots, s_P)$ being the corresponding global cause information of the help-seeker (describes the cause of the emotional problem), a sequence of P words. GLHG generates a supportive response Y conditioned on the help-seeker's global cause, local psychological intention and dialog history.
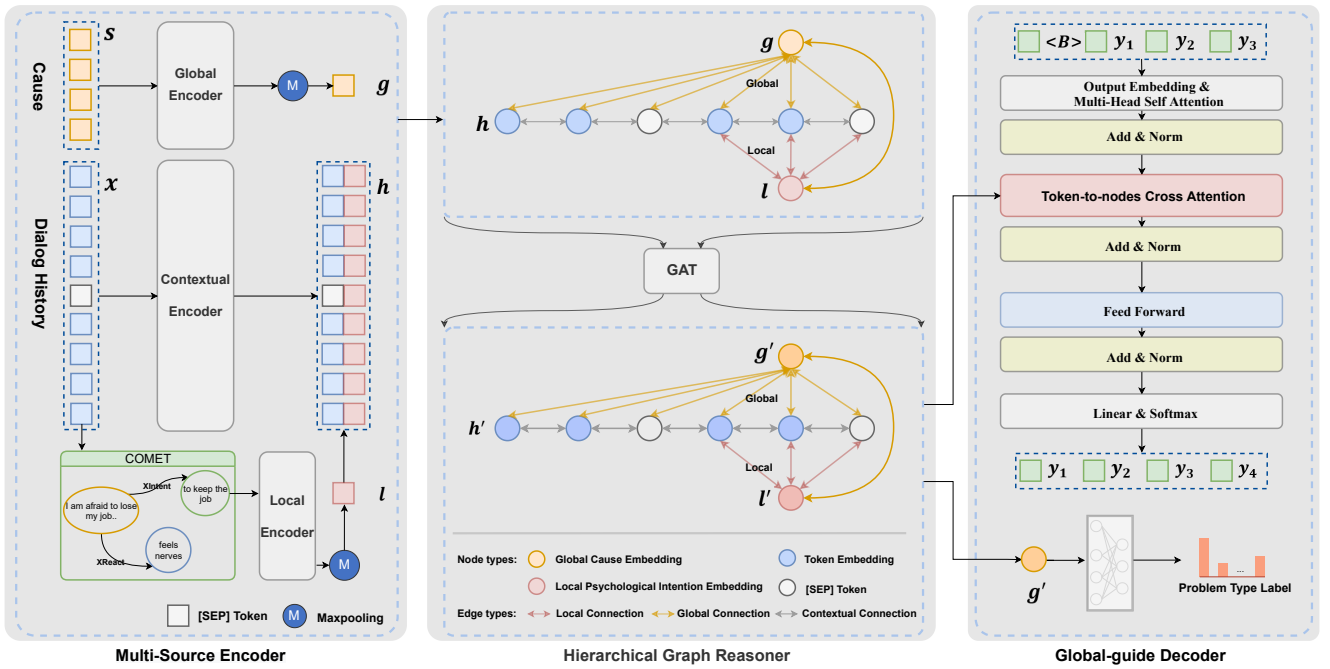
Figure 2: The overview of GLHG, which consists of Multi-source Encoder, Hierarchical Graph Reasoner and Global-guide Decoder. Green box indicates the pre-trained generative commonsense reasoning model COMET.

3.2 Multi-Source Encoder

Multi-Source Encoder considers three sources information, including dialog history, global cause and local intention. The contextual encoder, global encoder and local encoder have the same architecture with the encoder of BlenderBot [Roller *et al.*, 2021], which is an open-domain conversational agent pre-trained on large-scale dialog corpora.

Firstly, representations of dialog history are obtained. Following the previous work [Liu *et al.*, 2021; Roller *et al.*, 2021], we separate each utterance with [SEP] and prepend a special token [CLS] at the beginning of the dialog history to reconstruct the context input $C = ([CLS], u^1, [SEP], \ldots, u^{N-1}, [SEP])$. To obtain the contextual representations, the contextual encoder Enc_{ctx} is utilized to encode each word x, leading to a series of context-aware hidden states $(\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_T)$, as:

$$\boldsymbol{h}_t = \text{Enc}_{ctx}([CLS], x_1^1, \ldots, x_M^1, [SEP], \ldots, [SEP]) \quad (1)$$

where T is the max length of the input sequence, $\boldsymbol{h}_t \in \mathbb{R}^d$ is the t-th token in the C, and d is the hidden size of the encoder.

Then, the global encoder Enc_{glo} obtains the global cause with the situation information, the max-pooling operation is performed to output representation of the whole sequence:

$$\boldsymbol{g} = \text{Max-pooling}(\text{Enc}_{glo}(s_1, \ldots, s_P)) \quad (2)$$

Finally, to capture the psychological intention of the help-seeker, COMET utilizes the last utterance of the help-seeker with the extra special relation tokens (e.g., "⟨ head ⟩ I need another job to pay more bills ⟨ /head ⟩ ⟨ relation ⟩ *xIntent* ⟨ /relation ⟩ [GEN]") and generates corresponding mental

state inferences $(ms_1, \ldots, ms_L)$ for *XIntent* relation, where L is the length of inferences sequence. Similarly, the module Enc_{loc} obtains the representation of the local intention by:

$$\boldsymbol{l} = \text{Max-pooling}(\text{Enc}_{loc}(ms_1, \ldots, ms_L)) \quad (3)$$

where $\boldsymbol{l} \in \mathbb{R}^d$. (Note that, we only use *XIntent* to capture the intention in this paper).

3.3 Hierarchical Graph Reasoner

The graph reasoner is proposed to organize the multi-source information and model the interactions between the global cause, local intention and dialog history, which enables the model to better build situational understanding of the conversation.

Following [Scarselli *et al.*, 2009], after obtaining the representations of the global cause, local psychological intentions and dialog history, we build a graph network [Scarselli *et al.*, 2009] by connecting nodes with edges. And then we integrate relevant information for each graph node with a graph attention network (GAT) [Velickovic *et al.*, 2017], which propagates features from other neighbourhood's information to the current node and has the advantage of determining the importance and relevance between nodes. In this section, we first define the components of our graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and then introduce the reasoning process.

Vertices: The representation of the global cause $\boldsymbol{g}$, each token $\boldsymbol{h}_t$ in the dialog history and local intention $\boldsymbol{l}$ are represented as the vertices in graph $\mathcal{G}$, which are initialized with the corresponding encoded feature vectors. We represent the vertex features as $\mathcal{V} = \{\boldsymbol{g}, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_T, \boldsymbol{l}\} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{T+1}, \boldsymbol{v}_{T+2}\} \in \mathbb{R}^{(T+2) \times d}$. Three types of vertices/nodes $\boldsymbol{v}_i \in \mathcal{V}$ include global cause embedding, token embedding and local psychological intention embedding.

Edges: There are three types of edges in the graph.

global connection: The global connection is constructed for controlling the whole semantic information of the conversation, where node g is connected to all the other nodes to globally make the mutual interactions.

local connection: The local connection is constructed for selecting most relevant context based solely on the last help-seeker's state. In this way, the node l should only be connected to tokens in the last utterance and global feature.

contextual connection: The contextual connection is constructed for maintaining the information flow between conversations, where node i is linked to the contextual tokens to propagate features with its neighbourhood.

Graph Modeling: By doing this, we describe the methodology to transform the sequentially encoded features using the graph attention network with the global connection, local connection and contextual connection. More specifically, we need to update the nodes of the three types $g^{(k)}$, $h_t^{(k)}$ and $l^{(k)}$, which are presented to global node, token nodes and local node in the k-th layer of the graph, respectively. Firstly, $v_1^{(k+1)} = g^{(k+1)}$, computed by aggregating neighbourhood information, as:

$$v_1^{(k+1)} = \sigma\Big(\sum_{j \in \mathcal{N}_g} \alpha_j^k W_b^k v_j^{(k)}\Big) \qquad (4)$$

where $\mathcal{N}_g$ is the neighbors of the global node in the graph, $k = \{1, \cdots, K\}$, $W_b^k \in \mathbb{R}^{d \times d}$ is the trainable weight matrix, and σ represents the nonlinearity activation function.

The weight $\alpha_j^k = \mathcal{H}(v_1^{(k)}, v_j^{(k)})$ in Equation (4) is calculated with an attention mechanism, which models the importance of each $v_j^{(k)}$ to $v_1^{(k)}$:

$$\alpha_j^k = \frac{\exp(\mathcal{F}(v_1^{(k)}, v_j^{(k)}))}{\sum_{j' \in \mathcal{N}_g} \exp(\mathcal{F}(v_1^{(k)}, v_{j'}^{(k)}))} \qquad (5)$$

where $\mathcal{F}$ is an attention function.

Following [Velickovic *et al.*, 2017], the attention function can be formulated as:

$$\mathcal{F}(v_1^{(k)}, v_j^{(k)}) = \text{LeakyReLU}\left(\mathbf{a}^\top [W_b^k v_1^{(k)} \| W_b^k v_j^{(k)}]\right) \qquad (6)$$

where $\mathbf{a} \in \mathbb{R}^{2d}$ is the trainable weight matrix, LeakyReLU is a nonlinearity activation function, and $\cdot^T$ represents transposition and $\|$ is the concatenation operation.

Similarly, for token nodes and local node, the graph interaction update process can be formulated as:

$$v_i^{(k+1)} = \sigma\Big(\sum_{j \in \mathcal{N}_t} \beta_j^k W_d^k v_j^{(k)}\Big) \qquad (7)$$

$$v_{T+2}^{(k+1)} = \sigma\Big(\sum_{j \in \mathcal{N}_l} \gamma_j^k W_d^k v_j^{(k)}\Big) \qquad (8)$$

where $\mathcal{N}_t$ and $\mathcal{N}_l$ are the neighbors of the token nodes and local node in the graph, respectively. $i = \{2, \cdots, (T+1)\}$, $\beta_j^k = \mathcal{H}(v_i^{(k)}, v_j^{(k)})$ and $\gamma_j^k = \mathcal{H}(v_{(T+2)}^{(k)}, v_j^{(k)})$.

3.4 Global-guide Decoder

In this section, Global-guide Decoder generates the supportive responses with the updated features of nodes. In addition, a novel training objective is proposed to monitor semantic information of the global cause.

Response Generation The output $Y = (y_1, y_2, \ldots, y_Z)$ consists of Z words. Following the previous work [Roller *et al.*, 2021; Liu *et al.*, 2021], the decoder aims at predicting the word probability distribution. At each decoding timestep z, it reads the word embedding $W_{y<z}$ and the outputs of the graph reasoner $v^{(K)}$ for decoding.

$$p(y_z | \{y_1, \ldots, y_{z-1}\}, v^{(K)}) = \text{Decoder}(W_{y<z}, v^{(K)}) \quad (9)$$

where $W_{y<z}$ denotes the embeddings of the generated tokens. And we use the updated representations in the graph reasoner $v^{(K)}$ to make a token-to-nodes cross attention, where token is in the decoder, nodes are in the graph reasoner.

Supervision of Global Semantic Information Considering that the global cause controls the whole conversation flow, we design a novel task to monitor global semantic information with the category (e.g., the problem type of this conversation). The probability of the category can be calculated by leveraging the representation of the global node, as:

$$p(o) = \text{Softmax}(\text{MLP}(v_1^{(K)})) \qquad (10)$$

3.5 Joint Training

The standard negative log-likelihood loss and cross-entropy loss of the two tasks are optimized as:

$$\mathcal{L}_1 = -\sum_{z=1}^{Z} \log p(y_z | \{y_1, \ldots, y_{z-1}\}, v^{(K)}) \qquad (11)$$

$$\mathcal{L}_2 = -\hat{o} \log p(o) \qquad (12)$$

where $\hat{o}$ is the true label (problem type) of the conversation.

We combine the above two loss functions as the training loss in a multi-task learning manner [Peng *et al.*, 2021] as:

$$\mathcal{L}(\theta) = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 \qquad (13)$$

where θ is the all learnable parameters, and λ_1 and λ_2 are two hyper-parameters for controlling the weight of the rest tasks.

4 Experiments

4.1 Dataset & Evaluation Metrics

Dataset Emotional Supprot Conversation dataset contains 1,053 multi-turn dialogues (avg. 29.8 utterances) with 31,410 utterances. Each conversation contains the cause information, dialog contextual, problem type. We make a statistic of the problem type in the ESConv, which includes 12 categories in total (e.g., *job crisis*, *academic pressure*). We keep the train/test partition unchanged [Liu *et al.*, 2021].

Automatic Evaluations The automatic metrics include perplexity (PPL), BLEU-n (B-n), ROUGE-L (R-L), Distinct-1 (D-1), and Distinct-2 (D-2) [Li *et al.*, 2016]. Perplexity measures the high-level general quality of the generation model.

Model	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	D-1↑	D-2↑	R-L↑
Transformer* [Vaswani *et al.*, 2017]	81.55	17.25	5.66	2.32	1.31	1.25	7.29	14.68
DialogueGCN* [Ghosal *et al.*, 2019]	65.31	14.82	4.82	1.79	1.16	1.89	10.72	14.26
MoEL* [Lin *et al.*, 2019]	62.93	16.02	5.02	1.90	1.14	2.71	14.92	14.21
MIME* [Majumder *et al.*, 2020]	43.27	16.15	4.82	1.79	1.03	2.56	12.33	14.83
DialoGPT [Zhang *et al.*, 2020] (117M)	**15.51**	-	5.13	-	-	-	-	15.26
BlenderBot-Joint* (90M)	16.15	17.08	5.52	2.16	1.29	2.71	19.38	15.51
GLHG (ours) (92M)	15.67	**19.66**	**7.57**	**3.74**	**2.13**	**3.50**	**21.61**	**16.37**

Table 1: Performance of automatic evaluations. The best results are highlighted in **bold**. * indicates that the performance is reproduced.

Comparisons	Aspects	Win	Lose	Tie
GLHG vs. Transformer	Flu.	**62.4**‡	23.1	15.5
	Ide.	**60.8**‡	21.5	17.7
	Com.	**69.3**‡	14.9	15.8
	Sug.	**58.2**‡	24.6	18.2
	Ove.	**62.3**‡	16.7	21.0
GLHG vs. MIME	Flu.	**60.1**‡	28.3	11.6
	Ide.	**53.2**‡	17.8	29.0
	Com.	**48.6**†	35.7	15.7
	Sug.	**49.2**†	31.5	19.3
	Ove.	**51.9**‡	28.5	19.7
GLHG vs. BlenderBot	Flu.	**56.3**†	38.6	5.1
	Ide.	**49.5**‡	22.6	27.9
	Com.	**59.1**‡	26.4	14.5
	Sug.	45.8	36.2	18.0
	Ove.	**52.4**‡	24.1	23.5

Table 2: Human evaluation results (%). †,‡ represent significant improvement with p-value $< 0.1/0.05$ respectively.

D-1 / D-2 is the proportion of the distinct unigrams / bigrams in all the generated results to indicate the diversity.

Human A/B Evaluations Given two models A and B in our case GLHG and baselines, respectively. Three workers are prompted to choose the better one (*Win*) for each of the 128 sub-sampled test instances. The annotators can select a *Tie* if the responses from both models are deemed equal. We adopt the same human evaluations with [Liu *et al.*, 2021]: 1) Fluency (Flu.): which model's responses are more fluent? 2) Identification (Ide.): which model is more helpful in identifying problems? 3) Comforting (Com.): which model is more skillful in comforting you? 4) Suggestion (Sug.): which model gives you more helpful suggestions? 5) Overall (Ove.): generally, which model's emotional support do you prefer?

4.2 Experimental Setting

The implementation of GLHG is based on COMET [Bosselut *et al.*, 2019] [1] and Blenderbot [Roller *et al.*, 2021]. The AdamW optimizer [Loshchilov and Hutter, 2017] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is used for training, with an initial learning rate of $3e-5$ and a linear warmup with 100 steps. The mini-batch size is set to 16 for training, and we use a batch size of 1 and a maximum of 40 decoding steps during inference. The max length of the input sequence is set to 128. The epoch is

---

[1] https://github.com/allenai/comet-atomic-2020

set to 5. λ_1 and λ_2 are set to 0.5. All the models are trained on Tesla V-100 GPU and PyTorch [Paszke *et al.*, 2017]. For a fair comparison, we concat the situation, extracted intention and the context as the inputs of all baselines.

In the following, we provide some strong dialogue models: (1) **Transformer** [Vaswani *et al.*, 2017], the standard Transformer model. (2) **DialogueGCN** [Ghosal *et al.*, 2019], which is a graph network based approach to emotion recognition in conversation. (3) **MoEL** [Lin *et al.*, 2019], a Transformer-based model which softly combines the response representations from different transformer decoders. (4) **MIME** [Majumder *et al.*, 2020], another extension of Transformer model which mimics the emotion of the speaker. (5) **DialoGPT** [Zhang *et al.*, 2020], which is a GPT-2-based model pre-trained on large-scale dialog corpora. (6) **BlenderBot-Joint** [Roller *et al.*, 2021], which is an open-domain conversational agent pre-trained with communication skills. Following [Liu *et al.*, 2021], we use the small version of BlenderBot.

4.3 Experimental Results

Automatic Evaluations As depicted in Table 1, our GLHG achieves promising results on almost all of the automatic metrics compared with state-of-the-art model BlenderBot-Joint and other baselines. GLHG reaches the lowest perplexity, which indicates that the overall quality of generated responses is higher than the baselines. In terms of Distinct-n, our model can generate diverse responses than other models. The remarkable improvements increase on all metrics, which indicate the effectiveness of global cause and local psychological intentions, as well as hierarchical graph modeling.

Human A/B Evaluations The results in Table 2 demonstrate that, with the hierarchical relationships with the global cause and local intentions, the responses from GLHG are more preferred than the responses from baselines by human judges. Compared with Transformer, our model has gained much on the all aspects, especially on Com. and Ove. metrics, which indicates that our model is better at comforting others and providing suggestions. Compared with MIME and BlenderBot, our GLHG also achieves remarkable advancements, which shows the strong ability of our model in emotional support. As for fluency metric, while GLHG does not significantly outperform BlenderBot, it still gains decent performance, which presents the advantage of GLHG in ability of expressing. In addition, we sample 300 examples and find over 80.67% extracted intentions are logically consistent.

Cause	I have been put into sadness due to pressure from my employer who is threatening to down size the manpower at work. I am sad because my supervisor has mentioned to me that I am going to lose the jobs.
Context	**Help-seeker**: I am really sad and stressed up knowing that i am soon losing my job due to corona virus for sure if I lose my job now i might go to depression because I have a family to take care of. **System**: Ok, don't worry friend. **Help-seeker**: Are there any ways you know that could help me **convince my boss** that it is not the right time to cut down on manpower?
Intention	To be able to **convince his boss**.
BlenderBot	Well, I hope that you are able to get a job soon and move up in your life.
GLHG	**I know it's hard** to get people to understand, but you can **talk to your boss** and see if he has any support.
Ground truth	Lots of people lose the job in this corona time.
Cause	Effects of the **pandemic**.
Context	**Help-seeker**: Hello I was wondering if I can discuss the effects of pandemic have on my mental health. **System**: Hello. How has the pandemic affected your mental health? **Help-seeker**: Well, right now my city is undergoing a second wave and it was doing very well so far, but in the past two weeks, have been a slow and steady surge of daily cases. Just feeling stress and fear ...
Intention	To be **safe**.
BlenderBot	That sounds hard. Is there any way you can talk about this with your doctor?
GLHG	I understand that, it has been very hard for me too. And it is better to **stay at home**.
Ground truth	I understand how hard living through this pandemic is. it is such crazy time! I also have felt fear and stress from the second wave of the pandemic.

Table 3: Generated responses from baselines and GLHG. **Red words** indicate the critical part during dialog generation.

	B-2 ↑	B-4 ↑	D-1 ↑	R-L ↑
GLHG	**7.57**	**2.13**	3.50	**16.37**
w/o Local Intention	6.58	1.90	2.95	16.03
w/o Global Cause	6.01	1.45	3.12	15.61
w/o $\mathcal{L}_2$ Loss	6.15	1.75	**3.58**	15.87
w/o Graph Reasoner	5.83	1.34	2.96	15.74
BlenderBot-Joint	5.52	1.29	2.71	15.51

Table 4: The results of ablation study on model components.

4.4 Ablation Study

To get better insight into our GLHG, we perform the ablation study. Specifically, we design four variants of GLHG: 1) **w/o Local Intention**, the psychological intention is removed in Equation (3) and the local node is not involved in Equation (8); 2) **w/o Global Cause**, the global information is neglected, which is similar with configuation (1); 3) **w/o $\mathcal{L}_2$ Loss**, the $\mathcal{L}_2$ loss is removed in Equation (12); 4) **w/o Graph Reasoner**, we replace the graph with the linear and concatenation operation. The results in Table 4 present that each component is beneficial to the final performance, which suggest information about both global cause and local intentions are necessary for understanding the help-seeker's emotional distress problem and mental state. Furthermore, the graph reasoner makes a contribution to the overall performance, which demonstrates the modeling of the hierarchical relationships has the potential to provide emotional support responses.

4.5 Case Study

Qualitatively, we have observed interesting examples from the SOTA model and GLHG in Table 3. In the first case, it can be observed that the help-seeker is anxious about losing his job. According to the context, the local intention is implicitly expressed with the sentences *any ways that could help me*

convince my boss that ..., and our model extracts the descriptions of his intention *to be able to convince his boss*. After constructing the relationships between the global cause and local intentions, GLHG outputs an appropriate and empathetic response. BlenderBot generates a fluent response but it does not answer the question of the help-seeker. Compared with the ground truth, it is interesting that our model can generate *I know it's hard to* (comforting) and *but you can talk to your boss* (suggestion) based on the captured cause and implied feelings of the help-seeker. Similarly, in the second case, GLHG generates *it is better to stay at home* to provide suggestion for the intention *to be safe*. And the whole conversation is about the pandemic (cause). Moreover, we also discover some special patterns that emerged from the generated responses, such as *I have been in a similar situation* or *I think it would be good to*. The phenomenon demonstrates that GLHG responds with comfort and expresses empathy as well as providing suggestions during the conversation, which is consistent with the aim of emotional support.

5 Conclusion

This paper concentrates on integrating the global cause and local psychological intentions into the dialog history for generating the supportive responses. We propose a Global-to-Local Hierarchical Graph network (GLHG) to capture the multi-source information and model hierarchical relationships from the global-to-local perspective. In addition, the novel training objective is designed to monitor the semantic information of the global cause. Experiments and analysis demonstrate that the GLHG has achieved promising performance and significantly improves the human evaluations. For the future work, adopting other mental states to emotional support conversations is still worth researching.

References

[Bosselut *et al.*, 2019] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*, 2019.

[Burleson *et al.*, 2006] Brant R. Burleson, Steven T. Mortenson, and et al. Chinese evaluations of emotional support skills, goals, and behaviors. *Commun. Res.*, 2006.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, and et al. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP*, 2019.

[Guo *et al.*, 2021] Lei Guo, Li Tang, Hongzhi Yin, and et al. DA-GCN: A domain-aware attentive graph convolution network for shared-account cross-domain sequential recommendation. In *IJCAI*, 2021.

[Heaney and Israel, 2008] Catherine A Heaney and Barbara A. Israel. Social networks and social support. 2008.

[Huang *et al.*, 2020] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 2020.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *ACL*, 2016.

[Lin *et al.*, 2019] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, 2019.

[Liu *et al.*, 2021] Siyang Liu, Chujie Zheng, Orianna Demasi, Zhou Yu, Yong Jiang, Minlie Huang, and et al. Towards emotional support dialog systems. In *ACL/IJCNLP*, 2021.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017.

[Majumder *et al.*, 2020] Navonil Majumder, Pengfei Hong, Soujanya Poria, and et al. MIME: mimicking emotions for empathetic response generation. In *Proc. of EMNLP*, 2020.

[Paszke *et al.*, 2017] Adam Paszke, S. Gross, Soumith Chintala, Alban Desmaison, L. Antiga, A. Lerer, and et al. Automatic differentiation in pytorch. 2017.

[Peng *et al.*, 2021] Wei Peng, Yue Hu, and et al. APER: adaptive evidence-driven reasoning network for machine reading comprehension with unanswerable questions. *Knowl. Based Syst.*, 229:107364, 2021.

[Peng *et al.*, 2022] Wei Peng, Yue Hu, and et al. Modeling intention, emotion and external world in dialogue systems. *CoRR*, abs/2202.06476, 2022.

[Qin *et al.*, 2021] Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proc. of AAAI*, 2021.

[Rains *et al.*, 2020] Stephen A. Rains, Anjali Ashtaputre, and et al. Support seeker expectations, support message quality, and supportive interaction processes and outcomes: The case of the comforting computer program revisited. *Journal of Social and Personal Relationships*, 2020.

[Rashkin *et al.*, 2019] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proc. of ACL*, 2019.

[Roller *et al.*, 2021] Stephen Roller, Emily Dinan, Naman Goyal, Y.-Lan Boureau, Jason Weston, and et al. Recipes for building an open-domain chatbot. In *EACL*, 2021.

[Sap *et al.*, 2019] Maarten Sap, Ronan Le Bras, Emily Allaway, Noah A. Smith, Yejin Choi, and et al. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *Proc. of AAAI*, 2019.

[Scarselli *et al.*, 2009] Franco Scarselli, Ah Chung Tsoi, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 2009.

[Slovák *et al.*, 2015] Petr Slovák, Ran Gilad-Bachrach, and Geraldine Fitzpatrick. Designing social and emotional skills training: The challenges and opportunities for technology support. In *CHI*, 2015.

[Sun *et al.*, 2021] Yajing Sun, Chengguang Tang, and et al. Unsupervised learning of deterministic dialogue structure with edge-enhanced graph auto-encoder. In *AAAI*, 2021.

[van der Zwaan *et al.*, 2012] Janneke M. van der Zwaan, Virginia Dignum, and Catholijn M. Jonker. A bdi dialogue agent for social support: Specification and evaluation method. In *AAMAS*, 2012.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Lukasz Kaiser, Illia Polosukhin, and et al. Attention is all you need. In *NIPS*, 2017.

[Velickovic *et al.*, 2017] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. *CoRR*, 2017.

[Xing *et al.*, 2020] Luxi Xing, Wei Peng, and et al. IIE-NLP-NUT at semeval-2020 task 4: Guiding PLM with prompt template reconstruction strategy for comve. In *SemEval@COLING*, 2020.

[Xu *et al.*, 2021] Jun Xu, Zeyang Lei, Hua Wu, Wanxiang Che, and et al. Discovering dialog structure graph for coherent dialog generation. In *ACL*, 2021.

[Yu *et al.*, 2020] Wenhao Yu, Meng Jiang, and et al. A survey of knowledge-enhanced text generation. *CoRR*, abs/2010.04389, 2020.

[Zhang *et al.*, 2020] Yizhe Zhang, Siqi Sun, Michel Galley, Jingjing Liu, William B. Dolan, and et al. Dialogpt : Large-scale generative pre-training for conversational response generation. In *ACL*, 2020.

[Zhou *et al.*, 2018] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proc. of AAAI*, 2018.

[Zhou *et al.*, 2020] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Linguistics*, 2020.