Multi-Party Empathetic Dialogue Generation: A New Task for Dialog Systems

Lingyu Zhu[1], Zhengkun Zhang[1], Jun Wang[2]
Hongbin Wang[3], Haiying Wu[3], Zhenglu Yang[1]*
[1]TKLNDST, CS, Nankai University, [2]MS, Ludong University,
[3]Mashang Consumer Finance Co., Ltd., China
{zhulingyu,zhangzk2017,junwang}@mail.nankai.edu.cn,
{hongbin.wang03,haiying.wu02}@msxf.com,
yangzl@nankai.edu.cn

Abstract

Empathetic dialogue assembles emotion understanding, feeling projection, and appropriate response generation. Existing work for empathetic dialogue generation concentrates on the two-party conversation scenario. Multi-party dialogues, however, are pervasive in reality. Furthermore, emotion and sensibility are typically confused; a refined empathy analysis is needed for comprehending fragile and nuanced human feelings. We address these issues by proposing a novel task called Multi-Party Empathetic Dialogue Generation in this study. Additionally, a Static-Dynamic model for Multi-Party Empathetic Dialogue Generation, SDMPED, is introduced as a baseline by exploring the static sensibility and dynamic emotion for the multi-party empathetic dialogue learning, the aspects that help SDMPED achieve the state-of-the-art performance.

1 Introduction

Empathetic conversation studies have been coming to the forefront in recent years owing to the increasing interest in dialogue systems. Empathetic dialogues not only provide dialogue partners with highly relevant contents but also project their feelings and convey a special emotion, that is, empathy. As revealed by previous studies (Fraser et al., 2018; Zhou et al., 2020), empathy can enhance conversation quality and transmit appropriate emotional responses to partners. Accordingly, most, if not all, existing work focuses on taking an emotional perspective in dialogue studies (Levinson et al., 2000; Kim et al., 2004; Bertero et al., 2016; Fraser et al., 2018; Rashkin et al., 2019).

Although the empathetic conversation has received extensive attention, its exploration is still limited to the scenario with only two parties. In fact, multi-party chatting scenes are common in seminar discussions, conferences, and group chats.

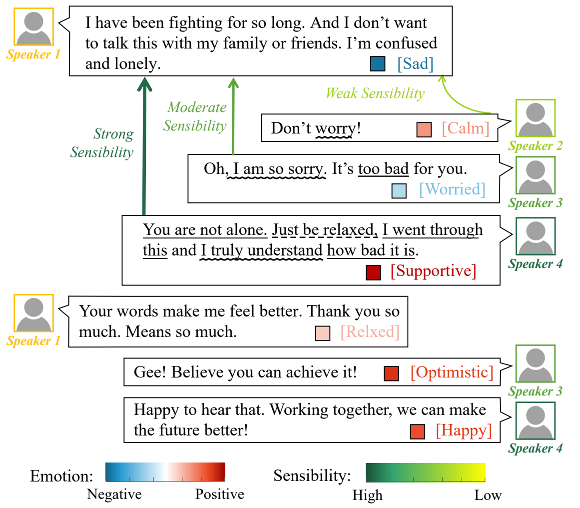

Figure 1: An empathetic dialogue example of multi-party. When people with different sensibilities respond to the same requests for help, their emotions and empathy differ. Different shades of red and blue denote the degree of positive and negative emotions, and different shades of green denote the degree of sensibilities. The texts use three kinds of underlines: straight, wavy, and dotted, which depict appropriate Emotional Reactions, Interpretations, and Explorations (three criteria to assess empathy), respectively.

Multi-party conversations also rely on aid from empathy analysis. For instance, people with a similar experience can smoothly communicate with each other and easily feel understood, encouraged, and supported. These observations encourage us to present a novel natural language processing task called Multi-Party Empathetic Dialogue Generation.

Generating multi-party empathetic dialogues faces two challenges. One challenge is the way to model multi-party dialogues. First, existing two-party dialogue models follow a seq2seq structure, whereas most multi-party dialogues are non-sequential. As shown in Figure 1, in response to *Speaker 1*, the third and fourth utterances both express empathy for his/her stress and struggle. Sec-

*Corresponding author.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 298 - 307
May 22-27, 2022 ©2022 Association for Computational Linguistics

ond, in addition to the target participant, other participants also have implicit influence and interaction, and should be considered of generating utterances at each step. For instance, as an example of how to successfully resolve the situation, *Speaker 4* inspires *Speaker 1* as well as relieves *Speaker 3* of his/her worry.

Another challenge is the way to model the fragile and nuanced feelings of dialogue participants. We first clarify the relations of sensibility, emotion, and empathy in this study. Previous empathy studies recognized the emotion of one party and generated dialogues coupled with the same emotion (Rashkin et al., 2019; Shin et al., 2020). However, empathy is also determined by sensibility, which is a perspective-taking ability to experience other partners' emotions and make an appropriate response with his/her own view. According to the response "I went through this" in Figure 1, we can find that *Speaker 4* has a similar experience to *Speaker 1*, while *Speaker 2* can only provide superficial comfort to *Speaker 1* due to his/her weak sensibility. We observe that sensibility arises from personality and experience, and remains static throughout a conversation. On the other hand, emotion may dynamically change. For example, *Speakers 2, 3*, and *4* possess different sensibilities to *Speaker 1*, and these personal background-related attributes are persistent in the conversation. By contrast, the emotion of *Speaker 1* gets reversed after receiving positive replies, as well as the main tone of this dialogue.

To comprehensively cope with the aforementioned challenges in this study, we present a **S**tatic-**D**ynamic model for **M**ulti-**P**arty **E**mpathetic **D**ialogue Generation called SDMPED. SDMPED models multi-party dialogues by constructing a dynamic graph network with temporal information and explores participants' dynamic emotions and static sensibilities by fusing speaker information.

The contributions of our work are as follows:

- We propose a new task called Multi-party Empathetic Dialogue Generation, which attempts to resolve the emotional changes and empathy generation of multiple participants in a conversation.

- We propose an effective baseline model SDMPED for this new task, which combines dynamic emotions and static sensibilities from multiple parties.

- We demonstrate that our approach leads to performance exceeding the state of the art when trained and evaluated on multi-party empathetic data.

2 Related Work

2.1 Empathy Analysis

Considering empathy in modeled conversations has been proposed as early as 20 years ago (Levinson et al., 2000). However, this idea has not been widely studied in NLP field due to the limitations of the available data. Recently, Rashkin et al. (2019) re-introduced the concept of empathetic dialogue and constructed the first empathetic dialogue dataset, EMPATHETICDIALOGUES (ED), which contains 32 emotions in 25K dialogues. Another dataset, PEC (Zhong et al., 2020), provides assurance that most of the data are in line with the characteristics of empathy, yet it lacks emotion-related annotations. Another limitation is that data in PEC come from only two forums on Reddit (i.e., happy5 and offmychest). The data in BlendedSkillTalk dataset (Smith et al., 2020) are collected from the ED, ConvAI2 (Dinan et al., 2020), and Persona-Chat (Zhang et al., 2018) datasets. However, only a small portion of these data are characterized by empathy. Notably, none of the aforementioned datasets have multiple (>2) persons participating in the same conversation, neither they include empathy degree labels.

Shin et al. (2020) formulated a reinforcement learning problem to maximize the user's emotional perception of the generated responses. Li et al. (2020b) utilized the coarse-grained dialogue-level and the fine-grained token-level emotions, which helped better capture the nuances of user emotions. In Caire (Lin et al., 2020), the empathy generation tasks are reinforced with an auxiliary objective for emotion classification by using a transfer learning model. Nevertheless, current empathetic dialogue models are conducted in the context of two participants; they do not explore the implicit interactions among multiple speaking persons and do not consider the differences in their sensibilities.

2.2 Multi-Party Dialogue

There have been quite a few studies on multi-party conversations before (Strauss and Minker, 2010), but they all focused on speech rather than conversational text. A recent multi-party study (Meng et al., 2018) has tended to focus on the Address and

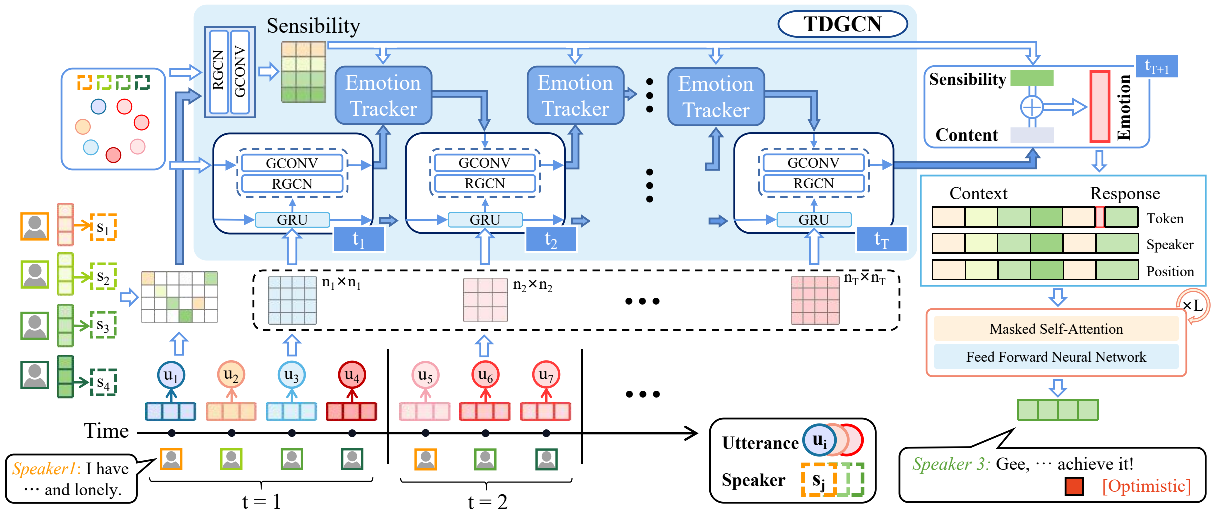

Figure 2: The overall architecture of SDMPED. Feature extraction provides the utterance and speaker sensibility nodes u_j and s_i, which will be input into TDGCN. By considering the utterance nodes and a segmented edge matrix E_t at time step t, we are able to compute the emotion-related content features. We combine static sensibilities with the current content information to get dynamic emotional information and input into the next moment. Finally, we use prompt tuning to generate final dialogue responses based on the dynamic emotions at $t + 1$.

Response Selection (ARS) task and ignore the influence of emotions, which is a significant departure from our empathetic dialogue task.

Over the last years, researchers have gradually shifted from studying simple emotions in two-party dialogues (Busso et al., 2008; Li et al., 2017) to conducting more complex emotion analysis of multiple participants. STAC (Asher et al., 2016) and ARS (Ouchi and Tsuboi, 2016) are the multi-party dialogue datasets without emotion labels. MELD (Poria et al., 2019) and MESID (Firdaus et al., 2020) create the multi-modal multi-party emotional dialogue datasets from the TV series *Friends*. However, these two datasets contain the emotion-related data derived from short and colloquial chats from TV series, and consequently, their dialogue quality cannot be guaranteed. Additionally, these datasets can only be utilized for simple upstream tasks, such as emotion recognition. Most of the dialogues in current datasets are daily conversations on trivial topics, while those modeling empathy dialogues are lacking.

Majumder et al. (2019) proposed a conversational emotion recognition model based on RNN to dynamically model the states of multiple speakers. Later, Ghosal et al. (2019) and Li et al. (2020a) also studied context and speaker sensitivity based on the approach of Majumder et al. (2019). A common problem of these models is that they only focus on the accuracy of emotion recognition while ignoring the dynamic changes of emotions.

3 Model

In this section, we introduce a static-dynamic model called SDMPED as shown in Figure 2. We begin by describing the construction of the Temporal Dynamic Graph Network (TDGCN), including speaker sensibility nodes, emotion-related utterance nodes, and various types of edges between them. Thereafter, we use TDGCN to obtain dynamic emotions and static speaker sensibilities by integrating nodes and edges. Finally, we use prompt tuning to generate final dialogue responses based on emotion and sensibility information.

3.1 Problem Definition

We regard an empathetic post and its meaningful replies as a dialogue and ensure that each dialogue has more than three participating speakers. A post contains replies from multiple people, along with associated emotion and empathy degree labels. The empathy degree label of each utterance will be used in conjunction with the emotional content in our future model to learn the sensibility of each person.

We propose a concept called dialogue emotional turn, which is different from the traditional dialogue turn. Specifically, a dialogue is assumed to have multiple sentences in one emotional turn but with the same emotional tone. When a person utters a second sentence, the emotion may already differ from the previous one. Other people's subsequent utterances and emotions will be centered around this sentence. Therefore, we divide the dialogues to study the emotion variations over time, according to the principle that the same speaker can make at most one utterance during each emotional turn.

Then, we introduce key symbols and concepts used in our study. A T emotional turns dialogue with N utterances between M ($M > 2$) speakers can be expressed as $U = \{u_{ik}|1 \leq i \leq N$ and $1 \leq k \leq M\}$, where u_{ik} represents the ith sentence from jth speaker. To better study emotion variations, we specify that a speaker can at most utter one sentence in each emotional turn. Thus, U can be divided into $U = \{U_t|1 \leq t \leq T\}$, where each part U_t has n_t nodes. Further, the sensibilities of speakers can be expressed as $S = \{s_1, s_2, ..., s_M\}$. Our model aims to generate an empathy response of length L.

3.2 Graph Construction

SDMPED captures the sensibility information and emotional variations of multiple parties owing to a novel graph network.

First, we train the multi-scale TextCNN (Zhang and Wallace, 2015) according to the empathy degrees of our dataset, and we extract the d-dimensional utterance-level features containing sensibility information. In each turn, we use the emotion of the first speaker as the main emotional tone, and extract the emotional content features based on those emotion labels in the same way.

Using these sensibility-related features as nodes and speaker-utterance relationships as an adjacency matrix, we construct a two-step static graph network to determine the static sensibility information $H_S = \{(H_x)_S|1 \leq x \leq M\}$ of speakers. Thereafter, we represent the dialogue as a directed graph $G = (V, E, R, W)$ to obtain additional emotional information. The graph is constructed as follows:

Nodes V: The node set $V = \{v_{ik}|1 \leq i \leq N$ and $1 \leq k \leq M\}$ incorporates emotion-related utterances. Among them, each node v_{ik} (abbreviated as v_i) is initialized with the extracted feature u_i spoken by the speaker s_k.

Adjacency Matrix E: E represents the adjacency matrix between emotion-related utterances. $e_{ij} \in E$ represents the edge from the utterance node v_i to v_j.

Edge Relations R: The relationship r_{ij} of edge e_{ij} is set mainly depending upon two things (Ghosal et al., 2019; Yang et al., 2021): the relative occurrence positions of u_i and u_j in the conversation (with three types of relations, namely, *Before*, *Current*, and *After*) and both speakers of the constituting utterance nodes, as shown in Figure 3.

Edge Weights W:Based on our assumptions, the edge weights are based on similarity-based attention, and the edge weights $\alpha_{ij} \in W$ are calculated as follows:

$$\begin{aligned} &\alpha_{ij} = \text{softmax}(u_i^T W[u_{i-p}, ..., u_{i+f}]), \\ &\text{for} \quad j = i - p, ..., i + f. \end{aligned} \tag{1}$$

And the relationship between the utterance and its speakers α_{ki} in static graph network can also be represented as $\frac{c}{Freq}$. Speaking frequency of the speakers $Freq$ denotes the utterance number of a speaker in the whole conversations. c is a speaking coefficient to avoid over-fitting.

Time Division Before feeding it into TDGCN, we need to divide E into T steps: $E = \{E_t|1 \leq t \leq T\}$. At time step t, the divided matrix E_t includes only edges corresponding to the utterance in the emotional turn t.

As shown in Figure 1, four speakers participate in the dialogue with 7 utterances. This dialogue has two emotional turns: u_1 to u_4 and u_5 to u_7. The nodes and edges are constructed in Figure 3. We take node u_3 as an example. The edge e_{13} represents that u_1 spoken by s_1 appears before u_3 spoken by s_3 and the influence between them; the self-loop e_{33} represents the influence of current node u_3 on itself.

Two-Step Graph Update: The graph update mechanism has been implemented in two steps in order to better track conversation information and dynamic emotions. The update mechanism is calculated as follows:

$$\begin{aligned} h_i^{(1)} &= \sigma(\sum_{r \in R} \sum_{j \in N_i^r} \frac{\alpha_{ij}}{c_{i,r}} W_r^{(1)} u_j + \alpha_{ii} W_0^{(1)} u_i), \\ h_i^{(2)} &= \sigma(\sum_{j \in N_i^r} W^{(2)} h_j^{(1)} + W_0^{(2)} h_i^{(1)}), \end{aligned} \tag{1}$$

where α_{ij} and α_{ii} are the edge weights and N_i^r denotes the neighboring indices of node v_i under relation $r \in R$. $c_{i,r}$ can be set in advance, such as $c_{i,r} = |N_i^r|$. σ is the activation function ReLU, while $W_r^{(1)}$, $W_0^{(1)}$, $W^{(2)}$, and $W_0^{(2)}$ are learnable parameters.

Utilizing the Two-Step Graph Update mechanism, we can effectively normalize the local neighborhood through neighborhood connections and enable self-dependent feature transformation through

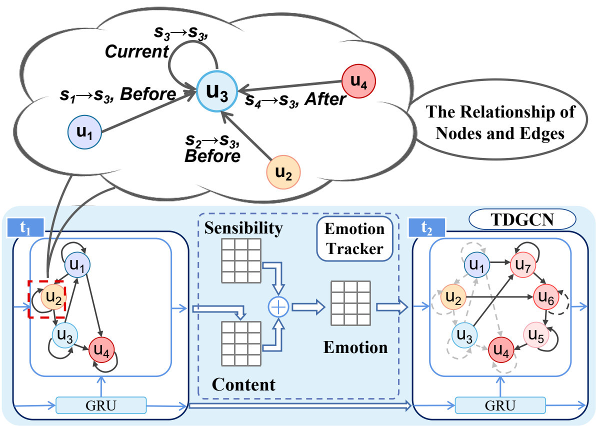

Figure 3: Transformation of dynamic emotions from t_1 to t_2, as well as various types of edges between different nodes (e.g., Node u_3).

self-connections, thereby extracting further information (Ghosal et al., 2019): We can call these two steps RGCONV and GCONV respectively in Figure 2.

3.3 TDGCN

Previous dynamic graphs were mostly used in spatio-temporal traffic networks with separated spatial and time features (Guo et al., 2019; Zhao et al., 2020). However, given that the utterance node is time-related and changes frequently, we implement the dynamic graph by updating a weight matrix through GRU and updating the hidden layer through the two-step graph:

$$
\begin{aligned}
M_t^{(l)} &= \mathrm{GRU}(H_{t-1}^{(l)}, M_{t-1}^{(l)}), \\
H_t^{(l)} &= \mathrm{GCONV}(\mathrm{RGCONV}(E_t, H_{t-1}^{(l)}, M_t^{(l)})),
\end{aligned} \tag{2}
$$

where $t \in [1, T]$ and $l \in [1, L]$ (L generally equals 2) denote the time and layer index, respectively. $M_{t-1}^{(l)}$ represents the weight matrix updated by GRU. $H_t^{(0)}$ is equal to the node features V. The hidden state $H_t^{(l)}$ of the lth layer at time step t can be divided into n_t parts: $H_t^{(l)} = \{(h_x)_t^{(l)}\}$, where x represents the speaker index. By concatenating person's sensibility with corresponding emotion-related content $(h_x)_t^{(l)}$, we obtain dynamic emotion embedding:

$$
(e_x)_t^{(l)} = \left[(H_x)_S; (h_x)_t^{(l)}\right]. \tag{3}
$$

Then, the emotion embedding set $e_t = \{(e_x)_t^{(l)}\}$ is sent to a fully connected layer and regarded as H_t at $t+1$ time step. We can also obtain a cross-entropy loss function at $t+1$:

$$
\begin{aligned}
P_e &= softmax(W_l e_{t+1}), \\
L_{emo} &= -\log\left(P_e[e]\right).
\end{aligned} \tag{4}
$$

3.4 Decoder and Loss

We adopt prompt tuning (Lester et al., 2021) to generate responses, which is a lightweight alternative to fine-tuning the generation task and keeps language model parameters unchanged while optimizing the prompt. The prompt adjustment achieves comparable performance in the full data setting by learning only parameters with a small proportion.

The representation e_{t+1} is first transformed by a linear transformation into prompt. We can obtain the input of the empathy decoder $Z = [X; \mathrm{prompt}; Y]$, where X and Y represent the context and target response, respectively. We use the standard maximum likelihood estimate to optimize the response prediction, and we obtain another loss function through the decoder:

$$
L_{res} = -\log\big(p(Y|R_{generate})\big). \tag{5}
$$

Finally, all the parameters are jointly trained end-to-end to optimize the listener selection and response generation by minimizing the sum of two losses:

$$
L = L_{emo} + L_{res}. \tag{6}
$$

4 Experiments

4.1 Dataset

Data Pre-Processing The MPED data is obtained from an online peer-to-peer support platform, where users can express their emotions by chatting with others who have similar experiences. Generally, we permit the words of each utterance to range between 3 and 100, excluding emojis, which are stored separately[1]. We discard artificially repeated characters, correct spelling errors, and standardize network language. Developing a dialogue model requires more ethical considerations. Therefore, we focus our analysis on help-seeking or emotional comfort-seeking conversations. As a result, the conversations with sensitive contents are filtered out. In the end, we further ensure that no private information is included.

[1]Emotional utterances have been incorporated in MPED yet not in our proposed baseline since we focus on unimodal text in this study.

It is quite beneficial that emotional category labels are available, which saves a lot of manual work. We have confirmed their accuracy and constructed the MPED dataset with kinds of emotions. We further classify these emotions for simplicity into 10 types, that is, *happy*, *sad*, *calm*, *angry*, *excited*, *exhausted*, *supportive*, *bored*, *nervous*, and *thankful*. MPED includes single-turn and multi-turn dialogue data, called MPED-S and MPED-M. We randomly split them into 80% training set, 10% validation set, and 10% testing set, respectively.

Empathetic Pre-Processing Given that empathy is a complex feeling, gathering empathetic data is challenging. We first remove the conversations that do not contain empathetic posts, such as games, and so forth. Then, we design a three-point scale (0 to 2) and evaluate empathy, where three criteria are used: Emotional Reactions (expressing warmth and compassion), Interpretation (articulating understanding of feelings and experiences), and Exploration (exploring feelings and experiences not stated in the post). Considering manually screening dialogues is infeasible on large-size data, we filter out simple replies and label single-turn dialogues. In the end, three degrees of empathy are included in MPED, that is, *weak*, *moderate*, and *strong*.

4.2 Experimental Setting

The hyper-parameters in our approach are set as follows. The input embeddings are 300-dimensional pre-trained 840B GloVe vectors. The speaking coefficient c is 5. The learning rate is 0.003 and batch size is 16. The dropout rate is 0.6, while the loss weight is $5e^{-4}$.

4.3 Evaluation Criteria

Automatic Evaluation Criteria We calculate the AVG BLEU (average of BLEU-1,-2,-3,-4) (Papineni et al., 2002) and ROUGE-L (Lin, 2004) scores as evaluations of model response generation, which have been often used to compare the system-generated response against the human-gold response in generation tasks.

Human Evaluation Criteria We randomly collect 100 dialogue samples and their corresponding generations from each model. Then, we assign human annotators to rate each response between 1 and 5 on three distinct attributes:

- *Empathy*: assesses whether the speaker of the response understands the feelings of others and fully manifests it;
- *Relevance*: evaluates whether the generated response is relevant with the dialogue context and consistent with the expressed information or background knowledge;
- *Fluency*: measures whether the response is smooth and grammatically correct.

4.4 Baselines and Models

MReCoSa: A context-sensitive model with multi-head self-attention (Zhang et al., 2019).
Multi-Trans: This multi-task model learns emotion classification and dialogue generation at the same time (Rashkin et al., 2018).
MoEL: This model (Lin et al., 2019) combines the response representations from multiple emotion-specific decoders.
EmpGD: This method (Li et al., 2020b) exploits coarse-grained and fine-grained emotions by an adversarial learning framework.
Caire: This method (Lin et al., 2020) fine-tunes a large-scale pre-trained language model with multiple objectives: response language modeling, response prediction, and dialogue emotion detection.
Random Prompt: We built a network with random values for prompt according to Lester et al. (2021).

We describe the variants of our model below:
Graph-Based: This simple model uses a graph-based model to build the empathetic dialogue graph of multi-party.
Two-Step Graph: This model adopts a graph network with two-step graph update.
SDMPED without Sensibility (SDMPED w/o S): This model ignores the sensibilities of speakers but maintains a TDGCN structure.
SDMPED: Our final model combines dynamic emotions with static sensibilities to produce empathy responses.

4.5 Experimental Results

Automatic Evaluation Results According to the experimental results shown in Table 1, our model SDMPED achieves the highest scores under most metrics compared with other baselines. The noticeable improvement indicates the effectiveness of SDMPED on empathetic expressions of multi-party. Since multi-party dialogues are not time-sequential

Model	MPED-M					MPED-S				
Metrics	ROUGE-L	AVG BLEU	Emp.	Rel.	Flu.	ROUGE-L	AVG BLEU	Emp.	Rel.	Flu.
MReCoSa	10.31	2.58	2.20	3.09	3.91	10.74	3.90	2.22	3.34	4.00
Multi-Trans	6.59	3.86	2.81	3.13	3.92	8.10	4.22	2.76	3.41	4.20
MoEL	6.83	2.99	3.11	3.07	3.89	8.44	3.13	3.00	3.28	4.13
EmpDG	10.86	4.26	3.19	3.39	4.30	11.53	4.52	3.32	3.55	4.30
Caire	11.58	4.85	3.17	3.62	4.37	12.48	5.49	3.30	**3.89**	4.46
Random prompt	11.36	4.68	3.10	3.65	4.10	12.04	5.41	3.44	3.81	4.40
SDMPED w/o S	12.06	5.57	3.29	3.66	4.30	13.47	5.88	3.51	3.81	4.53
SDMPED	**12.87**	**6.35**	**3.40**	**3.74**	**4.39**	**14.16**	**7.37**	**3.71**	3.86	**4.59**

Table 1: Experimental results on MPED. The automatic evaluations include AVG BLEU and ROUGE-L, and Emp.; Rel. and Flu. stand for the human evaluations *Empathy*, *Relevance* and *Fluency*.

Model	MPED-M		MPED-S	
Metrics	ROUGE-L	AVG BLEU	ROUGE-L	AVG BLEU
SDMPED	12.87	6.35	14.16	7.37
SDMPED w/o S	12.06	5.57	13.17	5.88
Two-Step Graph	11.54	4.87	12.39	5.69
Graph-Based	11.23	4.67	11.68	4.84

Table 2: Ablation study on MPED-M and MPED-S.

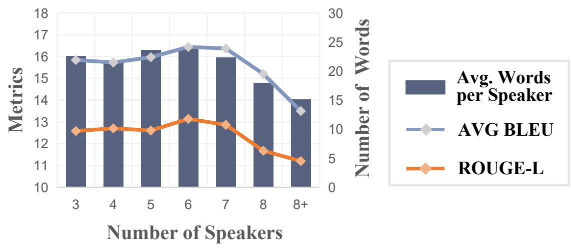

Figure 4: The effect of different numbers of speakers. The orange and blue lines represent BLEU-1 and ROUGE-L, and histograms in dark blue show the average number of words spoken by each person in multi-turn dialogues.

and multi-turn dialogues need to consider the impact of each turn, SDMPED performs better than the models MoEL, EmpDG, and Caire that are designed solely for two-party dialogue. Compared with the Random prompt model, our model has been greatly improved, which demonstrates that our emotional prompt design plays an important role. Given that persons have different sensibilities, adding the characteristics of different people to explore their conversations helps improve the performance. Thus, SDMPED obtains a performance improvement on the basis of SDMPED without Sensibility.

Human Evaluation Results Table 1 shows that SDMPED has achieved good performance in *Empathy*, *Relevance*, and *Fluency*. Our model is effective in capturing different emotional changes between multiple speakers and generating appropriate responses. MoEL and EmpDG are more inclined towards the characteristics of two-party dialogues, and thus cannot fully adapt to the new situation of multi-party. Random prompt and Caire are basically as good as our model in *Fluency*, however their *Empathy* and *Relevance* are inferior. These two models are pre-trained transfer learning models, and the generated responses are fluent and grammatical while being simple and general.

4.6 Ablation Study

We perform an ablation study to better understand the contributions of the main parts of our model. As shown in Table 2, the performance becomes noticeably worse, especially in the multi-turn dialogue data, after we remove the sensibility component. The degree of empathy for empathetic dialogues depends on the emotional tone at that time and the speakers' own abilities of perspective-taking, so studying sensibilities can help better investigate the responses generated by different people. According to the comparison of SDMPED without Sensibility and Two-Step Graph, emotions of people change at every moment, and updating the graph structure at each emotional turn is particularly necessary.

After removing the two-step graph update mechanism, we find that the results of Graph-Based have further declined, which indicates that the two-step graph convolution process can better extract empathetic and dialogue features.

4.7 Analysis of Speakers and Tokens

We investigate the effects of different numbers of speakers and tokens. When 3–7 speakers are available, as shown in Figure 4, the model maintains fairly stable results, indicating that it can handle multiple-party empathetic dialogues effectively. However, the results decline as the speaker number continues to increase. The reason for the drop is

	Speaker	Sensibility	Utterance
Context	*Speaker 1*	-	I am alone and have no friends now . I need a single hug . (Sad)
Response	*Speaker 2*	Weak	A virtual, because it could be possible. (Calm)
	Speaker 3	Moderate	**You are welcome to talk with me**. (Worried)
	Speaker 4	**Strong**	**I am sorry to hear that.** I believe **you can get through this** and **focus on what you love to do at the moment.** (Optimistic)
	Speaker 5	**Strong**	**Don't be miserable! Sending you sunshine to brighten your day**. (Supportive)

Table 3: An example of different responses by different speakers. Shades of blue represent the attention weights of *Speaker 1*. Below the text are three kinds of lines: straight, wavy, and dotted, which depict appropriate Emotional Reactions, Interpretations, and Explorations (three criteria to assess empathy).

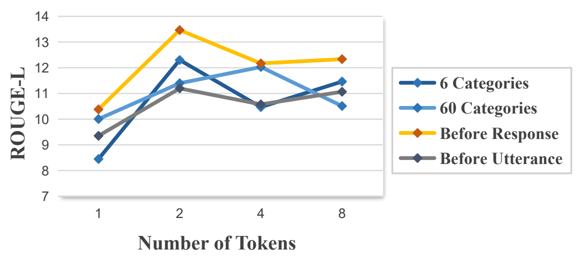

Figure 5: The effect of different numbers of tokens. The first three lines of this legend compare the effects when the emotion categories are 6, 10, and 60. **Before Utterance** and **Before Response** compare the effects of using different prompt embedding positions when dividing emotions into 10 categories.

that our conversations are typically concentrated between 3 to 5 people, and those with more than 7 people contain little content per speaker.

In Figure 5, we compare our model with two prompt embedding methods and different numbers of emotion classification categories. The comparison between the orange and blue curves shows that dividing emotions into 10 categories gives better results than the 6 and 60 categories (6 and 60 categories are similar to the number of categories in MELD and ED datasets). Clearly, dividing emotions into 10 categories and placing a prompt matrix with 2 tokens before the response can yield promising performance.

4.8 Case Study

We apply different speakers' sensibilities to the empathy decoder in the same multi-turn conversation context and obtain results based on MPED in Table 3. When presented with *Speaker 1*'s loneliness and depression, the following four speakers are willing to provide support, but they come up with different responses due to their different sensibilities. *Speaker 2* is relatively unable to appreciate the emotions of *Speaker 1* and jokes that he/she can find a virtual friend to hug; *Speaker 3* expresses warmth and *Speaker 4* and *Speaker 5* comfort *Speaker 1* and express their understanding. They also look forward to the future by suggesting that *Speaker 1* can do something that helps distract himself/herself.

5 Conclusions and Future Work

We have introduced a novel task called Multi-Party Empathetic Dialogue Generation.We have proposed a model called SDMPED suitable for the characteristics of the task. Our experiments have demonstrated that SDMPED is superior to other approaches on MPED. Future work can explore related issues such as integrating empathy into the dialogues, combining emojis and responses, guiding the active development of conversation.

Ethical Considerations

Data Collection. We collected publicly available data and removed all personal information (phone, email, postcode, location, and any other privacy information). Any potentially sensitive dialogues were completely removed from our data. No treatment recommendations or diagnostic claims were given in this study.

This research is approved and monitored by the University's Institutional Review Board and performed in accordance with the principle of GDPR (General Data Protection Regulation[2]) as follows: data processing shall be lawful if it is necessary for the performance of a task carried out in the public interest. Additionally, this study is explored not for any commercial use while merely for scientific

[2]https://gdpr-info.eu/.

purpose and public interest, which are safeguarded by the Art. 89 GDPR.

Annotator Compensation. We resorted to the Amazon Mechanical Turk crowdsourcing platform to evaluate three artificial indicators (i.e., Empathy, Relevance, and Fluency). The crowdworkers were assessed with 20 random sentences, which averagely took 5-6 minutes to accomplish, and compensated with $0.8 per HIT (Human Intelligence Task). The compensation was determined based on the US minimum wage of $7.12 per hour.

Potential Misuse. Our model is less likely to contribute to depression of users or generate non-empathic expressions (e.g., discrimination, criticism, and antagonism), since the model is based on the assumption that everyone has varying degrees of sensibility and empathy. Additionally, this model removes any sensitive information of users, and it is basically impossible to infer their personalities, preferences, interests, or other private information from the generated dialogues.

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China (No. 62106091) and Shandong Provincial Natural Science Foundation (No. ZR2021MF054).

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2721–2727.

Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. 2016. Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1047.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 335–359.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *Proceedings of the 33rd Conference on Neural Information Processing Systems Competition*, pages 187–208.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453.

Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. 2018. Spoken conversational AI in video games: Emotional dialogue management increases user engagement. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 179–184.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 154–164.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 922–929.

Sung Soo Kim, Stan Kaplowitz, and Mark V Johnston. 2004. The effects of physician empathy on patient satisfaction and compliance. *Evaluation & the Health Professions*, 27(3):237–251.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Wendy Levinson, Rita Gorawara-Bhat, and Jennifer Lamb. 2000. A study of patient clues and physician responses in primary care and surgical settings. *The Journal of the American Medical Association*, 284(8):1021–1027.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020a. Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200.

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

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 986–995.

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

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 121–132.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 09, pages 13622–13623.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive rnn for emotion detection in conversations. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pages 6818–6825.

Zhao Meng, Lili Mou, and Zhi Jin. 2018. Towards neural speaker modeling in multi-party conversation: The task, dataset, and models. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, volume 32.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BlEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

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

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

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7989–7993. IEEE.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030.

Petra-Maria Strauss and Wolfgang Minker. 2010. Dialogue management for a multi-party spoken dialogue system. pages 73–114.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1009–1021.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019. ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2020. T-GCN: A temporal graph convolutional network for traffic prediction. *IEEE Transactions on Intelligent Transportation Systems*, 21(9):3848–3858.

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

Li Zhou, Jianfeng Gao, Di Li, and Heung Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.