Towards Multi-Turn Empathetic Dialogs with Positive Emotion Elicitation

Anonymous ACL submission

Abstract

Emotional support is a crucial skill for many real-world scenarios, including caring for the elderly, mental health support, and customer service chats. This paper presents a novel task of empathetic dialog generation with positive emotion elicitation to promote users' positive emotion, similar to that of emotional support between humans. In this task, the agent conducts empathetic responses along with the target of eliciting the user's positive emotions in the multi-turn dialog. To facilitate the study of this task, we collect a large-scale emotional dialog dataset with positive emotion elicitation, called **PosEmoDial** (about 820k dialogs, 3M utterances). In these dialogs, the agent tries to guide the user from any possible initial emotional state, e.g., sadness, to a positive emotional state. Then we present a positive-emotion-guided dialog generation model with a novel loss function design. This loss function encourages the dialog model to not only elicit positive emotions from users but also ensure smooth emotional transitions along with the whole dialog. Finally, we establish benchmark results on PosEmoDial, and we will release this dataset and related source code to facilitate future studies.

1 Introduction

Emotion perception and expression are vital for building a human-like dialog system. Thanks to the availability of large-scale corpora and the rapid advances in deep learning, the potential of agents to improve the emotional well-being of users has been growing (Pamungkas, 2019, Huang et al., 2020). In particular, the agents could provide emotional support and prevention measures in against of the increasing stress level of individuals.

The previous researches on empathetic dialog generation, which focuses on conducting natural empathetic responding by understanding and acknowledging any implied feelings of users sheds light on enhancing user-agent emotional bond

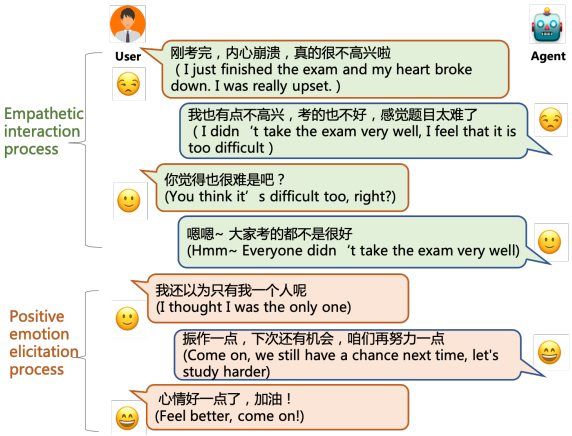

Figure 1: A sample of positive-emotion-guided empathetic conversation. It consists of two stages: (i) the agent expresses empathy about the situation of the user; (ii) the agent encourages the user and changes the emotion state of the user from "negative" to "positive".

(Rashkin et al., 2019, Li et al., 2020a). In Rashkin et al., 2019, a benchmark and dataset is proposed to make the dialogue system towards empathetic conversation. However, the user's emotional state at the end of the conversation are not sufficiently taken into account since current approaches only consider conducting empathetic responding in every turn of the dialogue. These models look backwards in the conversation context and might fail to jump out of user's negative emotion topics, limiting their applications in real-world scenarios, such as providing emotional support and caring for the elderly (Zhang and Danescu-Niculescu-Mizil, 2020).

Apart from that, positive emotion elicitation, which advance the conversation towards optimistic state to equip users to cope with the situation is also significantly related to positive outcomes of human interactions (Mishara et al., 2007, Sandoval et al., 2010, Lubis et al., 2019). Recently the studies (Lubis et al., 2017,Lubis et al., 2018, Li et al., 2020b) drew on an important potential of positive emotion elicitation in maximizing user emotional

Datasets	#Dialogs	Language	Emp.	P.E.G	Multi-turn	Source
NLPCC2017 (Huang et al., 2017)	1,119,207	Chinese	No	No	No	Weibo
MOJITALK (Zhou and Wang, 2018)	662,159	English	No	No	No	Twitter
PEC (Zhong et al., 2020b)	355,000	English	Yes	No	Yes	Reddit
Empatheticdialog (Rashkin et al., 2019)	24,850	English	Yes	No	Yes	Crowd Sourcing
DailyDialog (Li et al., 2017)	13,118	English	No	No	Yes	Online Websites
Enhanced SEMAINE (Lubis et al., 2018)	2,349	English	No	Yes	No	Crowd Sourcing
EmotionPush (Huang and Ku, 2018)	8,818	English	No	No	Yes	Facebook Message
MPDD (Chen et al., 2020)	4,142	English	No	No	Yes	TV-series
PosEmoDial (our dataset)	819,391	Chinese	Yes	Yes	Yes	Web

Table 1: Comparison of our dataset PosEmoDial with other datasets for emotional dialogs. Emp. denotes dialog empathy and P.E.G. denotes positive emotion guidance.

experience and promoting positive emotional states, similar to that of human beings. But these works usually attempt to conduct emotion elicitation in a single turn, yielding unnatural emotional transitions and thus failing to "reach an understanding" of the individuals with the absence of backwards empathetic reflection(Rogers and Carl, 2007, Hill and Nakayama, 2000, Lubis et al., 2017). Therefore, an ideal positive emotional elicitation process should progressively seek a certain degree of emotional resonance with the user (such as similar experiences, feelings) before improving user emotion towards a better state (Zhang and Danescu-Niculescu-Mizil, 2020). The multi-turn empathetic dialogs with positive emotion elicitation might yield mutually reinforcing advantages for agent's empathy and functionality of emotional support, which is less studied in previous work.

To sum up, we present a novel task, multi-turn empathetic dialog generation with positive emotion elicitation. In this task, the agent will first conduct empathetic responding and then naturally switch to positive emotion elicitation from users. Figure 1 provides an example for this task. To address this task, we encounter two challenges: (1) how to effectively capture emotions in an accurate and explainable way, (2) how to ensure smooth emotional transitions along with the whole dialog.

To facilitate the study of this task, we collect a human-to-human multi-turn Chinese dialog dataset with positive emotion elicitation (**PosEmoDial**). In PosEmoDial, every dialog is initiated by a speaker with either a positive, neutral, or negative emotion and ends up with a positive emotion of the same speaker that is elicited by another speaker. This dataset is collected from real web users in a web forum, not being annotated by crowdsourcing, which contains more natural dialog logic about how speakers successfully fulfill positive emotion elicitation (corresponding to *the second challenge*).

To address this task, we propose a novel Positive-emotion-guided empathetic dialog model (**PEGE**) by improving traditional negative log-likelihood (NLL) loss. Specifically, we introduce a new loss term, the Positive Emotion Guidance (**PEG**) loss, which measures how smoothly candidate responses at each dialog turn move from an initial emotion state at the first turn to the targeted positive emotion state at the last turn (corresponding to *the second challenge*). To enable PEG loss to measure the above emotional transitions more effectively, we employ an external resource, Valence-Arousal-Dominance (**VAD**) Lexicons (Mohammad, 2018), for representation of emotions in utterances (*the first challenge*). Our PEG loss encourages the dialog model to conduct positive emotion elicitation and also ensure smooth emotional transitions along with the whole dialog.

This work makes the following contributions:

- We present a novel task of empathetic dialog generation with positive emotion elicitation.
- We provide a large-scale empathetic dialog dataset with positive emotion elicitation, PosEmoDial.
- We propose a positive-emotion-guided pre-training-empowered dialog generation model (PEGE) with novel loss function design and confirm its effectiveness.

2 Related Work

Models for Emotional Dialogs Previous work on emotional dialogs fall into three categories: (1) controlled emotional dialog generation (Huang et al., 2018, Zhou et al., 2018, Colombo et al., 2019, Song et al., 2019, (Zhou and Wang, 2018),Shen and Feng, 2020, Zhong et al., 2020a); (2) empathetic dialog generation (Rashkin et al., 2019, Lin et al., 2019,

Context Emo	Negative	Neutral	Positive	Total
#Session	220,136	403,507	195,748	819,391
#Utterance	868,658	1,581,445	725,426	3,175,529

Table 2: Data scale of PosEmoDial, where Context Emo address the emotion of the first utterance by speaker. All sessions in PosEmoDial have at least three utterances (before deleting the last utterance), and the last utterance by user must be optimistic.

Majumder et al., 2020, Li et al., 2020a); (3) emotion elicitation (Lubis et al., 2018, Li et al., 2020b, Shin et al., 2019). Our model can conduct positive emotion elicitation, while previous work on empathetic dialog generation might fail to fulfill this dialog goal. Moreover, we emphasize natural emotional transitions through multi-turn dialogs, which is neglected by previous works on emotion elicitation.

Datasets for Emotional Dialogs To facilitate the study of emotional dialog, many researchers have created multiple datasets in previous works, as shown in Table 1. The two large-scale automatic annotated dataset NLPCC2017 (Zhou et al., 2018) and MOJITALK (Zhou and Wang, 2018) and the manually labeled dataset DailyDialog (Li et al., 2017) are widely used for controlled emotional dialog generation (Zhou et al., 2018, Zhou and Wang, 2018, Wang and Wan, 2019, Shen and Feng, 2020). The Empatheticdialog (Rashkin et al., 2019) dataset is designed for training empathetic dialog models (Lin et al., 2019, Majumder et al., 2020, Li et al., 2020a). The Enhanced SEMAINE dataset (Lubis et al., 2018) is constructed for the study of emotion elicitation by selecting or rewriting dialogs that can elicit positive emotion from SEMAINE corpus. In comparison with Empatheticdialog and Enhanced SEMAINE, our dataset is collected from dialogs between real web users, not through crowdsourcing. Then our dataset contains more natural emotional transitions logics with empathy and emotion elicitation naturally expressed. In addition, our dataset size is among the largest ones.

3 Dataset Construction

3.1 Task Definition

The person who starts the dialog is regarded as **user**, and the other one is regarded as **agent**. The goal of our task is to conduct empathetic dialog generation with positive emotion elicitation. There are two main characteristics of this task. Firstly, from the perspective of dialog goals, the agent should successfully elicit positive emotions from users through multi-turn dialogs. If the emotion state of users at the first dialog turn is negative or neutral, the agent should lead the dialog to a positive emotion state. If the initial one is positive, the agent should keep the emotion state to be positive or neutral. Secondly, from the perspective of emotional changes, the dialogue should be conducted in a natural, empathetic and gradual way.

3.2 Data Collection

In this work, we collect the dataset from natural dialogs of real web users on public websites, instead of through data annotation by crowdsourcing. The reason is that the empathy expressing of real users are more natural, and their chatting topics are more close to everyday life scenarios. We first collect Chinese dialogs from public social media and implement similar data cleaning process as done in Bao et al. (2020), which yielding a dataset containing 1.2 billion two-people dialog sessions. Then we introduce an ERNIE (Sun et al., 2019) based TextCNN (Kim, 2014) model to recognize the emotion of each utterance in dialogs. The detailed filtering procedures on the raw dataset are shown as follows:

1) The first utterance and the last utterances are from the same speaker who plays the role of user.

2) The probability of any negative or neutral or positive emotion in the first utterance is greater than 0.5. It helps us to improve the quality of emotion polarity information that is identified on this dataset.

3) The probability of any positive emotion in the last utterance is greater than 0.9. It also helps us to improve the quality of emotion related automatically-annotated information.

4) Delete dialogs with non-emotion related topics, such as renting, job hunting, blind date, which are not related to emotion eliciting but generally end up with positive utterance like "thanks" or "good" etc. (via keywords detection).

5) Delete dialogs with specific persons, institutions, address (being recognized with the use of Name Entity Recognition tools (Lample et al., 2016)) for privacy consideration.

6) Delete dialogs with offensive language (Kim, 2014) to decrease the probability of generating offensive responses.

Finally, we collect 819,391 dialogs that start with

any possible negative or neutral or positive emotion and end with a positive emotion, which we called PosEmoDial. Its statistics is provided in Table 2.

3.3 Data Processing

To learn how agent-side speakers conduct successful positive emotion elicitation, we delete the last utterance (from the user-side speaker) of each dialog, and require the model to predict agent-side response at each turn.

We denote the context as $\{u_1, ..., u_n\}$, the ground-truth response as r, the generated response as r'. For the sake of practicality, we treat the probability of the u_1 being emotionally positive $p(pos|u_1)$ or negative $p(neg|u_1)$ as the initial emotion state of the user-side speaker. For model training, we concatenate $p(pos|u_1)$ and $p(neg|u_1)$ with context and ground-truth response as the input.

4 Our Approach

The proposed model is based on PLATO-2 (Bao et al., 2020) where we only use the General Response Generation Stage[1] from PLATO-2 and improve its original loss function. The framework of our model is illustrated in Figure 2. Our proposed loss function consists of two components. The first one is traditional negative log-likelihood (NLL) loss. To effectively capture emotions in an accurate and explainable way and ensure smooth emotional transitions along with the whole dialog flow, we introduce two novel loss terms, the Positive Emotion Guidance (PEG) loss and Negative Emotion Regularization (NER) loss. The details of our model will be described in the followings.

4.1 Emotional Distance Calculation with VAD Lexicon

Previous works have shown the effectiveness of Valence-Arousal-Dominance (VAD) Lexicons for emotional dialog generation (Zhong et al., 2019, Colombo et al., 2019, Zhong et al., 2020a, Li et al., 2020a). We further validate the high accordance between VAD score and emotion polarity obtained by a well-trained ERNIE2-TextCNN emotion classifier (Sun et al., 2019, Kim, 2014). Therefore, for the sake of token-level generation control and model efficiency, the lexicon-based VAD vectors rather than neural network-based utterance representation is selected for emotion representation in our approach. We utilize the latest and largest VAD Lexicon, the NRC_VAD by Mohammad (2018), where Valence, Arousal, and Dominance are represented by continuous values in 0-1, indicating Negative to Positive, Claim to Excited, and Submissive to Dominant respectively. This lexicon includes 20,000 English vocabularies and their corresponding 13,870 distinct Chinese vocabularies. However, as there are 30k BPE tokens for the PLATO-2 lexicon. To fill this gap, we extends the NRC_VAD to cover all the PLATO-2 lexicon.

We define **Emotional Distance (ED)** as emotional changes across different utterances. Specifically, we employ the VAD lexicon to calculate the distance between the user initial emotion state and the generated response via a 2-Norm function, as shown in Eq.(1).

$$ED_t = \| \sum_{j=1}^{|u_1|} \frac{\mathbf{o_{u_{1,j}}}}{|u_1|} - \sum_{i=1}^{|V|} \mathbf{s_{t,i}} \mathbf{o_{w_i}} \|_2, \quad (1)$$

the first term calculates the expected VAD value of word in u_1, where $|u_1|$ denotes the length of the first utterance u_1, $u_{1,j}$ is the j-th word in u_1, $\mathbf{o_{u_{1,j}}} \in \mathbf{R^3}$ is a 3-dim vector representing emotion associated with the word $u_{1,j}$ in VAD space. The second term calculate the expected VAD value of the generated word at time step t, where $\mathbf{s_t} = \mathbf{softmax}(\mathbf{h_t})$ ($\mathbf{s_t} \in \mathbf{R}^{|\mathbf{V}|}$) is a confidence of the system of generating words $w_1, ..., w_{|V|}$ at time t. $\mathbf{o_{w_i}} \in \mathbf{R^3}$ is the VAD vector of i-th word in the vocabulary $[\mathbf{o_{w_1}}; ...; \mathbf{o_{w_{|V|}}}]$.

With the help of emotional distance, the new loss function is designed to ensure smooth emotional transitions along with the whole dialog flow as follows.

4.2 Positive Emotion Guidance Loss

The basic idea of the PEG loss is using emotional distance to control the emotion of response. The design process of PEG loss is described as follows:

1) If the user's starting emotion state is positive, the emotional state of the response is expected to align with the starting emotion state to keep the positive emotion of user in the whole dialogue. The PEG loss is designed as $\sum_{t=1}^{T} ED_t$, which will control the emotion of the response to be close to the starting emotion state, where ED_t is the measurement of emotional distance between the starting utterance and the generated response at time step t as illustrated in Eq.(1).

[1]There are two stages within the PLATO-2 model, the first stage conduct candidate responses generation and the second stage conduct responses selection. We only implement our work on the first stage of PLATO-2.

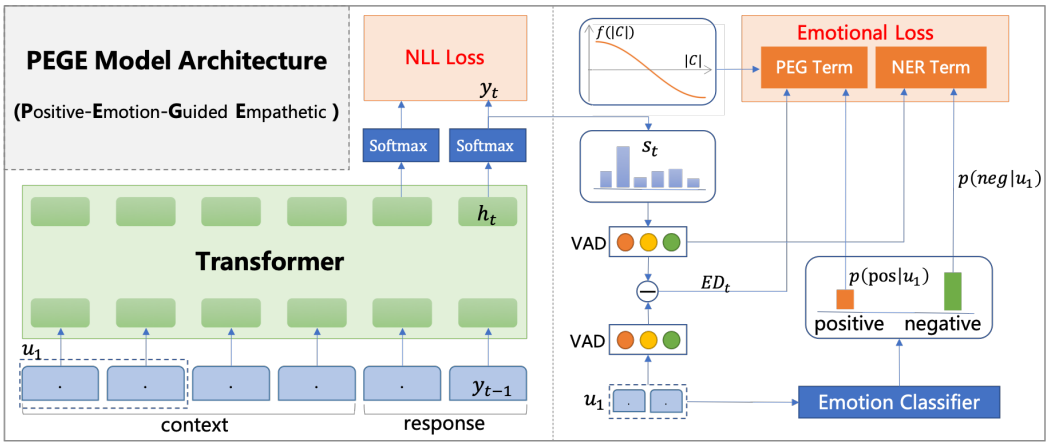

Figure 2: Illustration of our PEGE Model

2) If the user's starting emotion state is negative, the response is expected to express empathy at the dialogue's initial stage, and progressively transit to positive emotional state to elicit the user's positive emotion. Therefore, the emotional distance is required to be progressively increased throughout the whole dialog.

In order to progressively increase the emotional distance, we further improve the PEG loss by introducing a novel controlling function f, named as **Dialog Progress Function**. The $f(\cdot)$ is defined as:

$$f(|C|) = \cos(\frac{\pi}{max\_turn}|C|), \tag{2}$$

where max_turn is the maximum number of turns in dialog contexts, and it is set to 7 in our experiments. $|C|$ denotes the number of turns in dialog context at current time step. The $f(\cdot)$ value will transit from positive value to negative value as contexts get longer as shown in the middle part of Figure 2.

With the dialogue progress function, the PEG loss is redesigned as $\sum_{t=1}^{T}[f(|C|) \cdot ED_t]$. Then the emotion of the response will be controlled as follows:

- At the dialogue's initial stage, the emotional distance will be encouraged to be small. In other words, the emotion of response is controlled to align with the user's starting emotion to express empathy.

- At the dialogue's latter stage, the emotional distance will be encouraged to be big because of the negative value of function $f(|C|)$ results in negative loss. In other words, the emotion of response is controlled to be different from the starting emotion of user, which will be positive.

- At the whole dialogue stage, the emotional distance will be progressively increased from a small value to a big value because of the progressive transition of function $f(|C|)$. In other words, the emotion of response is controlled to express empathy at the dialogue's initial stage, and progressively transit to positive emotional state to elicit the user's positive emotion.

Finally, we use the probability of positive emotion of u_1 to combine the two kinds of the PEG loss as:

$$\begin{aligned} L_{peg} = \sum_{t=1}^{T}[&p(pos|u_1) \cdot ED_t \\ &+ (1 - p(pos|u_1)) \cdot f(|C|) \cdot ED_t], \end{aligned} \tag{3}$$

if a dialog starts with a positive emotion, $p(pos|u_1)$ will be close to 1, and the first term will play a leading role. If a dialog starts with a negative emotion, $p(pos|u_1)$ will be close to 0, and the second term will play a leading role. Otherwise, both will work.

4.3 Negative Emotion Regularization Loss

The potential drawback of the PEG loss is that the emotion of generated responses is required to align with u_1 at the initial stage. Therefore, the higher the probability of negative u_1 is, the more likely the PEG loss will encourage the generation of negative words at the initial dialog stage. Sometimes the responses containing these words can be injurious and offensive to users.

To address this issue, we add a NER loss to penalize the generation of too negative words with too small VAD values. The NER loss will be activated when u_1 is negative to balance the negative effect of the PEG loss. The NER loss is defined as:

$$L_{ner} = \sum_{t=1}^{T} p(neg|u_1) \cdot \| \sum_{i=1}^{|V|} \mathbf{s_{t,i}} \mathbf{o_{w_i}} \|_2, \quad (4)$$

where the notation is the same as described in the above PEG loss section.

4.4 Our Final Loss Function

The objective of the PEGE model is to minimize the following integrated Positive-emotion-guided Empathetic Loss (PEGE Loss) L_{pege}:

$$L_{pege} = L_{NLL}^{Baseline} + \alpha \cdot L_{peg} - \beta \cdot L_{ner}, \quad (5)$$

where $L_{NLL}^{Baseline}$ denotes the NLL loss:

$$L_{NLL}^{Baseline} = -\sum_{t=1}^{T} \log p(r_t|c, r'_{<t}), \quad (6)$$

where T is the length of the target response r and $r'_{<t}$ denotes previously generated words before time t.

The hyper parameter α and β in Eq.(5) denote the weights of PEG and NER loss respectively. We set $\alpha = 5$ and $\beta = 2$ for our final model based on grid search experiments.

5 Experiments

Following (Rashkin et al., 2019) , we conduct both automatic and human evaluations for dialog systems. Human evaluation is more convincing, as automatic metrics don't correlate well with human judgments of dialog quality (Liu et al., 2016).

5.1 Evaluation Metrics

Automatic evaluation metrics. Though BLEU and DISTINCT are two traditional metrics (Li et al., 2016, Lin et al., 2019), they have long been argued against its efficacy in open-domain dialogue generation (Liu et al., 2016), and either BLEU or DISTINCT is less relevant to our task. We keep them mostly as a reference.

To evaluate the efficacy of our model, we define three novel metrics that we describe next to account for the positive emotion guidance capability and emotion empathy capability.

PEG-Score: a new metric on a scale of [0,3] to measure the positive emotion guidance capability. It rewards the positive emotion the user obtained in the last half of utterances, i.e., $U_{user}^{last} = \{u_{-2}, u_{-4}, ..., u_{-n/2}\}$, and calculate the adjust averaged VAD values of each word in U_{user}^{last}. Sum up the averaged VAD values to obtain the PEG-Score:

$$PEG_{Score} = \sum_{VAD} \sum_{k \in U_{user}^{last}} \sum_{j=1}^{|u_k|} \frac{\mathbf{o_{u_{k,j}}} - \overline{\mathbf{o_{vad}}}}{|u_k|}, \quad (7)$$

E-Score: a new metric on a scale of [-3,0] to measure the emotion empathy capability. It penalizes the emotional distance between the agent responses and the user starting utterance (u_1) in the first half utterances, i.e., $U_{agent}^{first} = \{u_2, u_4, ..., u_{n/2}\}$, and calculates the averaged VAD values of each word in U_{agent}^{first}. We also calculate the averaged VAD for each word in u_1 as the starting emotion state. Then we subtract the two values and get their absolute VAD values. Sum up the absolute VAD values to obtain the E-Score:

$$E_{Score} = -\sum_{VAD} |\sum_{j=1}^{|u_1|} \frac{\mathbf{o_{u_{1,j}}}}{|u_1|} - \sum_{k \in U_{agent}^{first}} \sum_{j=1}^{|u_k|} \frac{\mathbf{o_{u_{k,j}}}}{|u_k|}|, \quad (8)$$

PEGE-Score: to balance the evaluation of positive emotion guidance and empathy, we sum up PEG-Score and E-Score to obtain the PEGE-Score (on a scale of [-3,3]):

$$PEGE_{Score} = PEG_{Score} + E_{Score}, \quad (9)$$

Human evaluation metrics. We run crowd-sourcing tasks at the level of both utterances and dialogs. Three crowd-sourcing workers are asked to score the response/dialog quality with a value of 0 or 1, and the final score is determined through the majority voting. These criterias are provided as follows:

Coherence: As an utterance level metric, it measures if the response is fluent, relevant and consistent with the context.

Informativeness: As an utterance level metric, it evaluates if the response is informative.

Positive emotion guidance: As a dialog level metric, it evaluates if the agent successfully guides the users from a non-positive emotion state to a positive emotion state, or keep their positive emotion state unchanged.

Empathy: As a dialog level metric, it is only measured when the positive emotion guidance score is 1 (else 0). It measures if the agent expresses empathy towards the user before positive emotion guidance, or keep the positive user not change as the criteria for positive emotion guidance.

5.2 Baselines

We select **MoEL** (Lin et al., 2019) and **MIME** (Majumder et al., 2020), two state-of-the-art baselines which solely introduce emotion as auxiliary information like our model in empathetic dialog generation tasks. **PLATO-2** (1.6B) (Bao et al., 2020) and **PLATO-2-FT** (fine-tuned version of PLATO-2 (1.6B) on PosEmoDial) which hold similar structure as our model are also selected.

However, since both MoEL and MIME are trained on the English dataset Empatheticdialog (Rashkin et al., 2019), we retrain them on PosEmoDial. For the sake of comparability, the semantic word embeddings of MoEL and MIME are initialized with the PLATO-2 embeddings (2048 dimensions).

5.3 Results

In multi-turn dialogue tasks, self-chat is a commonly used method to simulate human-bot conversations (Li et al., 2019, Roller et al., 2021), where a model plays the role of both partners in the conversation. For the sake of our task-specificity, we employ the original PLATO-2 model to play the role of the user. Because we want to simulate actual application scenarios as much as possible, a general "user" instead of an emotionally trained one is more appropriate. Accordingly, the candidate models will play the role of agent respectively.

The way to start the interactive conversation needs special attention. As pointed out by Roller et al. (2021) , if starting with 'Hi!', partners tend to greet with each other and only cover some shallow topics in the short conversation. Therefore, we construct 100 sentences as the starting utterance of different dialogues. Each sentence provides a specific context from the user's perspective, 33 of them are negative, 34 of them are neutral, and 33 of them are positive. The agent and "user" are required to perform self-chats given the context. There are 10 turns (20 utterances) in each dialog, including the input start utterance. We carry out automatic evaluation on the 100 self-chat logs and randomly select 50 conversations from 100 self-chat logs for human evaluation.

Automatic evaluation. Table 3 provides the automatic evaluation results for all the models. First, in terms of positive emotion elicitation, it shows that our model performs the best. Our model and PLATO-2-FT, which are fine-tuned on our PosEmoDial dataset, gain substantial improvements compared to PLATO-2. It indicates the effectiveness of our dataset for improving positive emotion elicitation capability. Moreover, when comparing our model with PLATO-2-FT, it can also be noted that the PEGE loss can provide an additional improvement on positive emotion guidance capability. Therefore, we conclude that our dataset and PEGE loss can work jointly to improve positive emotion guidance capability efficiently. Second, in terms of dialog empathy, our model gains the best performance as well. Our model's significant advantage over the second-best model PLATO-2-FT verifies the effectiveness of our loss design towards empathy capability. MoEL and MIME, which are not pre-trained on the large-scale corpus, are less capable of generating appropriate responses, hurting their empathetic dialog capability and resulting in a slightly worse E-Score than PLATO-2 and PLATO-2-FT. These results confirm the efficiency of our model in positive emotion elicitation while ensuring dialog empathy.

Human evaluation. Table 4 provides the human evaluation results for all the models. Our model has significantly better performance on two task-specific metrics (positive emotion guidance and empathy), considerably better performance on the coherence metric, and comparable performance on the informativeness metric. By comparing our model with PLATO-2-FT, our model obtains around 52% improvements on P.E.G. and 63% improvements on Emp. This remarkable result demonstrates the effectiveness of our PEGE loss on positive emotion guidance and empathy capability. Our dataset PosEmoDial also shows its effectiveness in training emotional dialog model as PLATO-2-FT fine-tuned on PosEmoDial outperforms PLATO-2 with 44% improvements on P.E.G. and 46% improvements on Emp. By applying PEGE loss and PosEmoDial simultaneously, our model gains 119% improvements on P.E.G. and 138% improvements on Emp. over PLATO-2, which further verifies the mutual benefits of our PEGE loss and PosEmoDial dataset.

Moreover, the models which get better performance on human evaluation metrics P.E.G. and Emp. also get higher scores on automatic evaluation metrics, PEG-Score, E-Score, and PEGE-Score. This result indicates the reliability of our proposed automatic metrics. We also observe that 81.37% of dialogues that successfully guide the user towards positive emotion express empathy be-

	Static Eval				Interactive Eval		
Models	BLEU1↑	BLEU2↑	Distinct-1↑	Distinct-2↑	PEG-Score↑	E-Score↑	PEGE-Score↑
MoEL	5.901%	2.077%	6.087%	19.728%	0.063	-0.214	-0.151
MIME	6.458%	2.117%	6.709%	19.372%	0.077	-0.202	-0.125
PLATO-2	**7.204%**	1.966%	8.418%	34.249%	-0.012	-0.189	-0.201
PLATO-2-FT	7.024%	**2.131%**	12.937%	44.512%	0.090	-0.185	-0.095
Ours	6.870%	2.039%	**13.266%**	**47.249%**	**0.160**	**-0.126**	**0.034**

Table 3: Comparison of automatic evaluation metric results under a static 5k test set and interactive self-chat dialogs among our model and baselines.

Models	Coh.↑	Inf.↑	P.E.G.↑	Emp.↑
MoEL	0.190	0.904	0.260	0.260
MIME	0.228	0.892	0.300	0.140
PLATO-2	0.934	**0.974**	0.320	0.260
PLATO-2-FT	0.916	0.954	0.460	0.380
Ours	**0.946**	0.962	**0.700**	**0.620**

Table 4: Comparison of human evaluation metric results on self-chat dialogs among our model and baselines. Coh., Inf., P.E.G. and Prog. stand for Coherence, Informativeness, Positive emotion guidance, and Empathy, respectively.

Models	PEG-Score↑	E-Score↑	PEGE-Score↑
$L_{NLL}^{Baseline}$	0.090	-0.185	-0.095
L_{ner}	0.068	-0.177	-0.109
L_{peg}	0.065	-0.134	-0.069
D_{plato}	-0.011	-0.191	-0.202
D_{pege}	0.072	-0.139	-0.063
Ours	**0.160**	**-0.126**	**0.034**

Table 5: Comparison of automatic evaluation metric results under interactive self-chat dialogues among our model, ablation models, and models on random dataset.

fore emotion elicitation. It verifies our proposed dialog task's rationality, i.e., expressing empathy before transit to positive emotion elicitation is crucial for building a human-like dialog system with emotion perception and expression capability.

5.4 Ablation Study

To evaluate the effect of the PEG loss and NER loss, we delete them respectively or simultaneously to get L_{ner}, L_{peg} and $L_{NLL}^{Baseline}$. We also eliminate the impact of PoSEmoDial by fine-tuning PLATO-2 and our model on 1M randomly selected dataset, denote as D_{plato} and D_{pege}. Note that when $L_{NLL}^{Baseline}$ is applied, the model is equivalent to the settings of PLATO-2-FT.

Table 5 illustrates the results of the ablation study. Our model with PEGE loss gets the best performance, and the model with L_{ner} gets bad performance on all metrics. With only NER loss (L_{ner}) remains, the model is more inclined to generate positive responses directly instead of conditioned on the user emotion state transition, which may not necessarily lead to positive feedback from users. This result is consistent with our real-world observations that the response to a negative statement with positive emotion directly without any emotional transition sometimes is inappropriate and even offensive. As the PEG loss L_{peg} is designed with both positive emotion elicitation capability and empathy capability, L_{peg} performs much better. However, without NER loss, the model with L_{peg} will endure the risk of generating excessively negative responses, which may sometimes be unacceptable to users as well, and therefore bring no gain with positive emotion elicitation. The results suggest that all components in PEGE loss L_{pege} are valuable and indispensable.

The comparison between D_{plato} and D_{pege} illustrates that our model is not data-dependent and can be generalized in other datasets since considerable improvements can be obtained on all three metrics even PEGE model is trained on randomly selected data. Meanwhile, PosEmoDial can actually facilitate model performance for both PLATO-2 and PEGE, validating its effectiveness in our task.

6 Conclusion

In this paper, we propose a novel task of multi-turn empathetic dialogs with positive emotion elicitation and collect a human-to-human Chinese multi-turn emotional dialog dataset with positive emotion elicitation (PosEmoDial). Then we propose a novel positive-emotion-guided empathetic dialog model (PEGE) by improving traditional NLL loss. The updated loss can encourage the dialog model to not only elicit positive emotions from users, but also ensure smooth emotional transitions along with the whole dialog flow. The results of the experiments confirm the usability of our dataset and the effectiveness of our model. In the future, we will introduce psychology-related domain knowledge to facilitate the modeling of in-depth emotional dialogs to support emotional counseling.

7 Ethical Considerations

We are sure that PosEmoDial has been collected in a manner that is consistent with the terms of use of any sources and the intellectual property and privacy rights of the original authors of the texts. Meanwhile, our project is approved by an IRB. Finally, we also provide details on the characteristics of PosEmoDial and steps taken to ensure the potential problems with the quality of the dataset do not create additional risks in Section 3.

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. PLATO-2: towards building an open-domain chatbot via curriculum learning. *CoRR*, abs/2006.16779.

Yi-Ting Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. MPDD: A multi-party dialogue dataset for analysis of emotions and interpersonal relationships. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 610–614. European Language Resources Association.

Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3734–3743. Association for Computational Linguistics.

C. E. Hill and E. Y. Nakayama. 2000. Client-centered therapy: Where has it been and where is it going? a comment on hathaway (1948). *Journal of Clinical Psychology*, 56(7):861–875.

Chenyang Huang, Osmar R. Zaïane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 49–54. Association for Computational Linguistics.

Chieh-Yang Huang and Lun-Wei Ku. 2018. Emotionpush: Emotion and response time prediction towards human-like chatbots. In *IEEE Global Communications Conference, GLOBECOM 2018, Abu Dhabi, United Arab Emirates, December 9-13, 2018*, pages 206–212. IEEE.

Minlie Huang, Zuoxian Ye, and Hao Zhou. 2017. Overview of the NLPCC 2017 shared task: Emotion generation challenge. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, volume 10619 of *Lecture Notes in Computer Science*, pages 926–936. Springer.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Trans. Inf. Syst.*, 38(3):21:1–21:32.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087.

Qintong Li, Piji Li, Zhumin Chen1, and Zhaochun Ren. 2020a. Towards empathetic dialogue generation over multi-type knowledge.

Shifeng Li, Shi Feng, Daling Wang, Kaisong Song, Yifei Zhang, and Weichao Wang. 2020b. Emoelicitor: An open domain response generation model with user emotional reaction awareness. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3637–3643. ijcai.org.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

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, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2017. Eliciting positive emotional impact in dialogue response selection. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems, IWSDS 2017, Farmington, PA, USA, 6-9 June 2017, revised selected papers*, volume 510 of *Lecture Notes in Electrical Engineering*, pages 135–148. Springer.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5293–5300. AAAI Press.

Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2019. Positive emotion elicitation in chat-based dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4):866–877.

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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8968–8979. Association for Computational Linguistics.

Mishara, L. Brian, Chagnon, François, Daigle, Marc, Balan, Bogdan, Raymond, and Sylvaine. 2007. Which helper behaviors and intervention styles are related to better short-term outcomes in telephone crisis intervention? results from a silent monitoring study of calls to the u.s. 1-800-suicide network. *Suicide & Life-Threatening Behavior*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20, 000 english words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 174–184. Association for Computational Linguistics.

Endang Wahyu Pamungkas. 2019. Emotionally-aware chatbots: A survey. *CoRR*, abs/1906.09774.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Rogers and R. Carl. 2007. The necessary and sufficient conditions of therapeutic personality change. *J Consulting Psychology*, 21(3):240.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 300–325. Association for Computational Linguistics.

J. Sandoval, A. N. Scott, and I. Padilla. 2010. Crisis counseling: An overview. *Psychology in the Schools*, 46(3).

Lei Shen and Yang Feng. 2020. CDL: curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 556–566. Association for Computational Linguistics.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *CoRR*, abs/1906.08487.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3685–3695. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Ke Wang and Xiaojun Wan. 2019. Automatic generation of sentimental texts via mixture adversarial networks. *Artif. Intell.*, 275:540–558.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, pages 5276–5289. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2020a. CARE: commonsense-aware emotional response generation with latent concepts. *CoRR*, abs/2012.08377.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7492–7500. AAAI Press.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020b. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6556–6566. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,(AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

Xianda Zhou and William Yang Wang. 2018. Mojitalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1128–1137. Association for Computational Linguistics.