Emp-RFT: Empathetic Response Generation via Recognizing Feature Transitions between Utterances

Wongyu Kim[1], **Youbin Ahn**[2], **Donghyun Kim**[2], **and Kyong-Ho Lee**[2]
[1]Department of Artificial Intelligence, [2]Department of Computer Science
Yonsei University, Seoul, Republic of Korea
{rladnjsrb9999, ybahn, dhkim92, khlee98}@yonsei.ac.kr

Abstract

Each utterance in multi-turn empathetic dialogues has features such as emotion, keywords, and utterance-level meaning. Feature transitions between utterances occur naturally. However, existing approaches fail to perceive the transitions because they extract features for the context at the coarse-grained level. To solve the above issue, we propose a novel approach of recognizing feature transitions between utterances, which helps understand the dialogue flow and better grasp the features of utterance that needs attention. Also, we introduce a response generation strategy to help focus on emotion and keywords related to appropriate features when generating responses. Experimental results show that our approach outperforms baselines and especially, achieves significant improvements on multi-turn dialogues.

1 Introduction

Humans have empathy which is the ability to understand situations others have experienced and emotions they have felt from the situations (Eisenberg and Strayer, 1987). That ability also enables to interest and console others while sharing a conversation. Thus, empathetic response generation task has been considered noteworthy. Figure 1 shows an example of a multi-turn empathetic dialogue dataset, EmpatheticDialogues (Rashkin et al., 2019) constructed to solve the task. A speaker talks about one of 32 emotion labels and a situation related to the emotion label, and a listener empathizes, responding to the speaker. Existing approaches (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Li et al., 2020; Kim et al., 2021) for the task achieve promising results but show limitations when dialogues become long because they extract features from the concatenation of all tokens in the context at the coarse-grained level.

However, at the fine-grained level, each utterance in multi-turn empathetic dialogues has features such as emotion, keywords that each denote

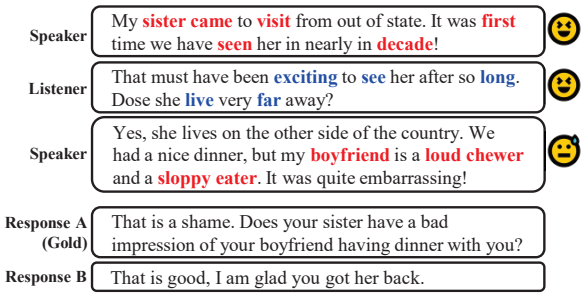

Figure 1: An example of EmpatheticDialogues with response A and B. Response B is from one of state-of-the-art models. Highlighted words are keywords.

what an interlocutor feels and primarily says, and utterance-level meaning that can be known when looking at the entire utterance. In addition, it is a natural phenomenon that features of each utterance differ from the previous, as the dialogue is prolonged. Hence, we humans instinctively recognize these feature transitions, which helps us understand how the dialogue flows and grasp the features of utterance that needs attention. Also, humans respond to others, focusing on emotion and keywords related to appropriate features. Take the example in Figure 1. In the first turn, the speaker is excited to see the speaker's sister in a long time by mentioning keywords (e.g., 'sister', 'visit', 'decade'), and the listener reacts to the excitement and asks about her by mentioning keywords (e.g., 'exciting', 'see', 'live'). However, in the second speaker utterance, the speaker becomes embarrassed because of the speaker's boyfriend's bad table manners by mentioning keywords (e.g., 'boyfriend', 'loud', 'eater'). We humans recognize that the features of second speaker utterance have changed compared to those of previous utterances, and usually decide to be attentive to the features of the second utterance. Then, by focusing on information such as keywords of that utterance and emotion and keywords (e.g., 'bad', 'impression') related to the features of that utterance, humans generate empathetic, coherent,

and non-generic responses like response A. However, the model which produces non-empathetic and incoherent response like response B, considers that the features of the first speaker utterance represent the context from the coarse-grained view.

In this paper, we first propose to annotate features on each utterance at the fine-grained level (§4). Then, we introduce a novel **Emp**athetic response generator based on **R**ecognizing **F**eature **T**ransitions (Emp-RFT), which has two essential parts: Feature Transition Recognizer and Response Generation Strategy. The first part recognizes feature transitions between utterances, utilizing comparison functions of Wang and Jiang (2017), which makes Emp-RFT understand the dialogue flow and grasp appropriate features of utterance that needs attention. The second part helps Emp-RFT focus on emotion and keywords related to appropriate features. Specifically, by fusing context with keywords, such keywords are emphasized within each utterance and get more attention when generating responses. Then, Emp-RFT detects next emotion and keywords that denote emotion and keywords of the response, which helps figure out proper emotion and keywords for generation. Lastly, inspired by Dathathri et al. (2020); Chen et al. (2020), a new mechanism of Plug and Play Language Model(PPLM), contrastive PPLM using contrastive loss, is introduced, which controls Emp-RFT to actively use the keywords detected to be next keywords when generating responses.

We conduct experiments on EmpatheticDialogues. Emp-RFT outperforms strong baselines, particularly, when dialogues are multi-turn.

Our main contributions are as follows. (1) We introduce a novel approach that recognizes feature transitions between utterances, which results in understanding how the dialogue flows and grasping the features of utterance that the model should be attentive to. (2) We propose a response generation strategy including fusing context with keywords, next emotion and keywords detection, and contrastive PPLM. The strategy makes our model focus on emotion and keywords related to appropriate features when generating responses. (3) In the experiments, Emp-RFT outperforms baselines, especially, when dialogues are prolonged.

2 Related Work

Since Rashkin et al. (2019) release EmpatheticDialogues, many approaches have been proposed to generate empathetic responses. Lin et al. (2019) propose mixture of emotional experts. Majumder et al. (2020) propose emotion grouping, emotion mimicry, and stochastic sampling. Li et al. (2020) extract emotional words through lexicon and propose an adversarial generative model. Shen et al. (2021) apply dual-learning with unpaired data for the bidirectional empathy. Gao et al. (2021) integrate emotion cause into response generation process through gated mechanism. Sabour et al. (2021); Li et al. (2022) use implicit commonsense for context modelling. Kim et al. (2021) train a model to extract words that cause the speaker's emotion and attach RSA Framework (Frank and Goodman, 2012) to any generative models to generate responses, focusing on emotion cause words.

Recently, many studies have shown remarkable improvements through recognizing transitions of features between utterances in open-domain multi-turn dialogues. Qiu et al. (2020) perceive transitions of emotion states for context modelling. Zou et al. (2021) propose a module to manage keyword transitions. Zhan et al. (2021) model external knowledge transitions to select a knowledge used for generation. In multi-turn empathetic dialogues, we consider emotions, keywords, and utterance-level meaning (Gu et al., 2021) as important features of each utterance and propose a novel approach of recognizing feature transitions between utterances.

3 Task Formulation

Given context $con = [u^1, \ldots, u^{n-1}]$, where an utterance $u^i = [u^i_1, \ldots, u^i_{|u^i|}]$ consists of $|u^i|$ words, we can obtain $e = [e^1, \ldots, e^{n-1}]$ and $k = [k^1, ..., k^{n-1}]$, where e^i and $k^i = [k^i_1, ..., k^i_{|k^i|}]$ each denote emotion and $|k^i|$ keywords of u^i through data preparation (§4). To conduct next keywords detection, we construct keyword pairs kps (§4.2) whose each pair has two keywords each from keywords of the speaker utterance and keywords of the listener utterance in the same turn. Finally, given con, e, k, and kps, we detect next emotion e^y and next keywords $k^y = [k^y_1, \ldots, k^y_{|k^y|}]$, and generate an empathetic response $y = [y_1, \ldots, y_m]$.

4 Data Preparataion

In this section, we introduce feature annotation in the speaker and listener utterances.

Feature	Top-1 Acc	Top-5 Acc	macro-F1
EofSU	46.77	81.26	43.55
EofLU	58.44	89.96	53.25
Feature	**TL-P**	**TL-R**	**TL-F1**
KofSU	41.53	66.58	51.15
KofLU	52.31	60.97	56.30

Table 1: Performances of feature annotations. When we evaluate annotations of **EofSU/KofSU**, EMOCAUSE (Kim et al., 2021) made based on EmpatheticDialogues, is used to verify emotion and emotion cause words detection models. For evaluations of keyword annotations, we use the metrics, Token-Level(**TL**) Precision(**P**), Recall(**R**), and **F1** (DeYoung et al., 2020), usually used in token extraction tasks.

4.1 Feature Annotation in Speaker Utterances

Emotion and Keywords of Speaker Utterance (EofSU/KofSU). Speakers try to say an emotional experience that causes a certain emotion in the utterance. Thus, we leverage a model (Kim et al., 2021) which is trained to jointly detect an emotion and emotion cause words of the speaker utterance, using EmpatheticDialogues. We regard top-6 emotion cause words as keywords and remove stopwords and punctuations in keywords.

4.2 Feature Annotation in Listener Utterances

Emotion of Listener Utterance (EofLU). We fine-tune RoBERTa (Liu et al., 2019) to detect an emotion given a situation description in EmpatheticDialogues. Then, the model predicts an emotion of the listener utterance.

Keywords of Listener Utterance (KofLU). Listeners express empathy in the utterance through three Communication Mechanisms (CMs) (Sharma et al., 2020) including emotional reaction, interpretation, and exploration. Thus, three models are leveraged, where each model is trained to detect words that cause one of three CMs, using another dialogue dataset for mental health support [1]. Then, three models predict such words in the listener utterance. Since predicted words take up slightly a lot in the listener utterance, these words are filtered out in the keyword pairs construction.

Keyword Pairs Construction. Inspired by Zou et al. (2021), keyword pairs kps are constructed not only to filter out above predicted words, but also to conduct next keyword detection. Given a dialogue

[1]A dialogue has a (post, response) pair, and words which cause each CM are annotated on each dialogue.

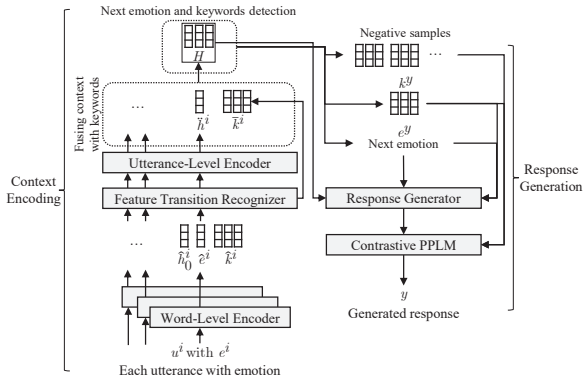

Figure 2: Overall architecture of Emp-RFT.

corpus, all pairs are extracted, where each pair has a head word and a tail word each from keywords in the speaker utterance and predicted words in the listener utterance in the same turn. Then, all pairs are filtered out to obtain high-frequency pairs through pointwise mutual information (PMI)[2] (Church and Hanks, 1990) which can measure the association between two words in a corpus. Filtered pairs become kps. A tail word of a kp is regarded as a keyword of the listener utterances joined to extract that keyword pair.

Performances of feature annotations are summarized in Table 1 and show reliable results. However, test sets for **KofLU** based on EmpatheticDialogues, don't exist. Thus, we randomly sample 100 test dialogues in EmpatheticDialogues and ask 3 human workers to annotate whether each word plays important role for empathizing in the listener utterances. By majority voting, the final verdict on each annotation is decided. We compute the inter-annotator agreement on annotation of test sets for **KofLU** through Fleiss' kappa ($\boldsymbol{\kappa}$) (Fleiss and Cohen, 1973), and result in 0.55, where $0.4 < \boldsymbol{\kappa} < 0.6$ indicates moderate agreement.

5 The Emp-RFT Model

In this section, we detail Emp-RFT whose overall architecture is shown in Figure 2.

5.1 Context Encoding

Word-Level Encoder. Emp-RFT contains an encoder $f_\theta(\cdot)$ which has the six-layer encoder of BART (Lewis et al., 2020) as the backbone and extracts feature vectors of each u^i. Inspired by BERT (Devlin et al., 2019), we prefix each utterance with a $[SEN]$ token, so $u_0^i = [SEN]$. Then, each token

[2]We use pairs whose PMI ≥ 1. The pairs whose tail words are stopwords or punctuations, are removed.

is represented as emb^i_j, the sum of the following four embeddings: word embedding, position embedding, role embedding and emotion embedding $M_e \in \mathbb{R}^{n_{emo} \times d}$ [3]. Then, the encoder transforms each utterance into a list of output hidden states:

$$[\hat{h}^i_0, ..., \hat{h}^i_{|u^i|}] = f_\theta([emb^i_0, ..., emb^i_{|u^i|}]), \quad (1)$$

where $\hat{h}^i_j \in \mathbb{R}^d$. For each utterance, we can obtain utterance-level meaning vector $\hat{h}^i_0$ derived from the token $[SEN]$, concatenated keyword vectors $\hat{k}^i \in \mathbb{R}^{|k^i| \times d}$ derived from the tokens corresponding to k^i_p (p is the index for keywords.), and emotion vector $\hat{e}^i = M_e \hat{h}^i_0$.

Feature Transition Recognizer. Emp-RFT has a component that operates as the process illustrated in Figure 3. The component computes feature transition information between feature vectors, utilizing two comparison functions, subtraction and multiplication of Wang and Jiang (2017). Each feature vector is compared to previous two feature vectors [4]. First, emotion transition information eti^i is computed:

$$eti^i = \mathrm{ReLU}(W_{eti}(f_{com}(\hat{e}^i, \hat{e}^{i-1}, \hat{e}^{i-2}))), \quad (2)$$

$$f_{com}(\hat{e}^i, \hat{e}^{i-1}, \hat{e}^{i-2}) = \begin{bmatrix} (\hat{e}^i - \hat{e}^{i-1}) \odot (\hat{e}^i - \hat{e}^{i-1}) \\ \hat{e}^i \odot \hat{e}^{i-1} \\ (\hat{e}^i - \hat{e}^{i-2}) \odot (\hat{e}^i - \hat{e}^{i-2}) \\ \hat{e}^i \odot \hat{e}^{i-2} \end{bmatrix}, \quad (3)$$

where f_{com} and $\odot$ each denote our transition information computing function and Hadamar product, and $W_{eti} \in \mathbb{R}^{d \times 4n_{emo}}$. Next, utterance-level meaning transition information uti^i is computed:

$$uti^i = \mathrm{ReLU}(W_{uti}(f_{com}(\hat{h}^i_0, \hat{h}^{i-1}_0, \hat{h}^{i-2}_0))), \quad (4)$$

where $W_{uti} \in \mathbb{R}^{d \times 4d}$. We then obtain enhanced utterance vector of each utterance by integrating utterance-level meaning vector, and emotion and utterance-level meaning transition information:

$$\bar{h}^i = \mathrm{FC}_{utt}([\hat{h}^i_0; eti^i; uti^i]), \quad (5)$$

where FC_{utt} is a fully-connected layer with size of d. In addition, keyword transition information

[3] j, d, and n_{emo} each denote the index for words, hidden size, and the number of emotion classes. The role and emotion embbedings are each for distinguishing two interlocutors and for incorporating the emotion into each utterance.

[4] If there aren't previous feature vectors, we can obtain those by regarding first output hidden state of a padded utterance as utterance and keyword vectors (Qiu et al., 2020).

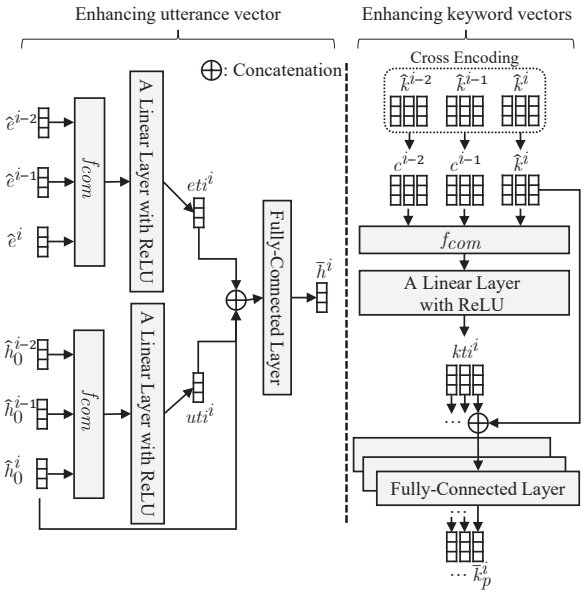

Figure 3: Operation process of feature transition recognizer.

kti^i is computed between concatenated keyword vectors and cross-encoded vectors c^t, where $t \in \{i-1, i-2\}$:

$$kti^i = \mathrm{ReLU}(W_{kti}(f_{com}(\hat{k}^i, c^{i-1}, c^{i-2}))^T), \quad (6)$$

$$c^t = \mathrm{softmax}(Q^i(K^t)^T)\hat{k}^t, \quad (7)$$

$$Q^i = \hat{k}^i W_Q, K^t = \hat{k}^t W_K, \quad (8)$$

where $W_{kti} \in \mathbb{R}^{d \times 4d}$, W_Q and $W_K \in \mathbb{R}^{d \times d}$. We can obtain enhanced keyword vector of each keyword by integrating keyword vector, and keyword transition information:

$$\bar{k}^i_p = \mathrm{FC}_{key}([\hat{k}^i_p; kti^i_p]), \quad (9)$$

where FC_{key} is a fully-connected layer with size of d. Consequently, the enhanced feature vectors guide Emp-RFT to accurately grasp the features of utterance that the model should be attentive to when given feature transition information.

Utterance-Level Encoder. Emp-RFT contains another encoder $g_\phi(\cdot)$ which has the six-layer encoder of BART, and transforms enhanced utterance vectors with global position embeddings (GPE) into a context representation to capture relationships between utterances (Gu et al., 2021):

$$[\ddot{h}^1, ..., \ddot{h}^{n-1}] = g_\phi([\bar{h}^1, ..., \bar{h}^{n-1}]). \quad (10)$$

Emp-RFT consists of hierarchical structures of encoders through word-level and utterance-level encoders. This structure makes Emp-RFT comprehend each utterance at the fine-grained level, and

understand the context by integrating information based on comprehension of each utterance.

Fusing Context with Keywords. Emp-RFT fuses context with keywords as the process illustrated in Figure 4. We first dynamically build keyword graph for each context. Keywords in each context become nodes and are initialized by corresponding enhanced keyword vectors with GPE. Edges are built across the below cases: (1) between two keywords from the same utterance and (2) between a keyword from a certain utterance and another keyword from the previous two utterances. Also, a tail word in a kp whose head word is k_p^{n-1} is appended as a node and connected with k_p^{n-1} node. Appended nodes (ANs) are initialized through BART decoder whose parameters are frozen with GPE, and used for next keywords detection. To obtain keyword representation $\hat{v}_o^i$ from the keyword graph(o is the index for nodes.), nodes are updated based on multi-head graph-attention mechanism (Veličković et al., 2018; Li et al., 2022). This mechanism makes Emp-RFT not only capture relationships between nodes but also manage influences of each appended node through attention architecture:

$$\hat{v}_o^i = v_o^i + \mathop{\Big\|}_{mh=1}^{MH} \sum_{z \in A_o^i} \alpha_{oz}^{i,mh} (W_v^{mh} v_z), \quad (11)$$

$$\alpha_{oz}^{i,mh} = \frac{\exp((W_q^{mh} v_o^i)^T W_{key}^{mh} v_z)}{\sum_{s \in \mathcal{A}_o^i} \exp((W_q^{mh} v_o^i)^T W_{key}^{mh} v_s)}, \quad (12)$$

where v_o^i, $\|$, A_o^i, and $\alpha_{oz}^{i,mh}$ each denote a node representation, the concatenation of MH attention heads, the neighbours of v_o^i in the adjacency matrix A, and self-attention weight and W_v^{mh}, W_q^{mh}, $W_{key}^{mh} \in \mathbb{R}^{d_{mh} \times d}$ ($d_{mh} = d/MH$). Lastly, we can obtain the fused context representation $H = [h^1, ..., h^{n-1}]$ by fusing the context representation with the sum of keyword representations:

$$h^i = \mathrm{FC}_{fuse}([\ddot{h}^i; \mathrm{sum}([\hat{v}_1^i, ..., \hat{v}_{|k^i|}^i])]), \quad (13)$$

where FC_{fuse} is a fully-connected layer with size of d. Consequently, keywords are emphasized within each utterance and get greater attention when generating responses.

Next Emotion and Keywords Detection. Emp-RFT detects next emotion e^y and keywords k^y, which helps figure out proper emotion and keywords for generation. First, based on the max-pooled fused context representation, next emotion

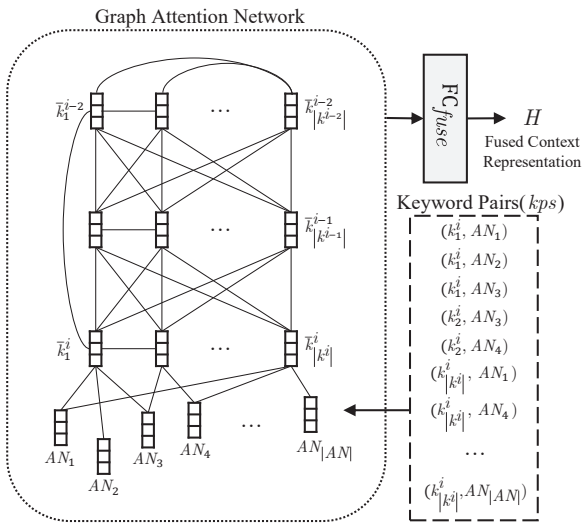

Figure 4: Operation process of fusing context with keywords. AN_o is the appended node. Some symbols and edges are omitted for simplicity.

distribution is predicted:

$$P_e = \mathrm{softmax}(M_e \mathrm{MP}(H)), \quad (14)$$

where MP denotes maxpooling. We use the emotion with the highest probability ($\hat{e}^y$) for generation. Also, Emp-RFT predicts whether the word of each AN belongs to the next keywords through the binary classification, where the true label denotes the word belongs to:

$$P_k = \prod_{o=1}^{|ANs|} \mathrm{softmax}(W_{AN}[\hat{v}_o^n; \mathrm{MP}(H)]), \quad (15)$$

where $W_{AN} \in \mathbb{R}^{2 \times 2d}$. We consider the words of ANs whose probabilities for the true label ≥ 0.8 as the keywords ($\hat{k}^y$) for generation.

5.2 Response Generation

Response Generator. Emp-RFT includes a response generator (RG) which has the six-layer decoder of BART as the backbone. Through the four embeddings with $\hat{e}^y$, explained previously, we can obtain the input sequence embedding for RG. We prefix it with the sum of node representations corresponding to $\hat{k}^y$. Then, RG is fed to predict probability distribution on each next token y_t based on the fused context representation:

$$P(y|con, e, k, kps) = \prod_{t=1}^{m} P(y_t|y_{<t}, H). \quad (16)$$

Training. We apply cross-entropy loss to three objectives (eq. 14, 15, 16), and train parameters of

Emp-RFT in end-to-end manner through the sum of all losses .

Contrastive PPLM. Analysis on the generated responses of the trained Emp-RFT shows that an active reflection of $\hat{k}^y$ is demanded. Thus, inspired by Dathathri et al. (2020); Chen et al. (2020), we propose Contrastive PPLM with a discriminator using contrastive loss. Existing discriminators (Dathathri et al., 2020; Majumder et al., 2021) are trained to predict whether a sentence contains a certain attribute, using cross-entropy loss. Then, the gradient of the loss is passed to the generative model to generate a sentence containing such attribute during inference. However, since keywords are not attributes but objects, we train a discriminator to predict whether a response in EmpatheticDialogues is more similar to the keyword set of the response(positive sample) than the keyword sets of another responses(negative samples) in the same batch, using contrastive loss based on the similarity between objects:

$$L^a_{pplm} = -\log \frac{\exp(r_a^T ks_a/\tau)}{\sum_{b=1}^{B} \exp(r_a^T ks_b/\tau)}, \qquad (17)$$

where r,ks,τ and B each denote response and keyword set representations, a temperature parameter and batchsize. During inference, we repeatedly sample three random ANs except for nodes of $\hat{k}^y$, and consider the sum of such AN representations as one of negative samples and the sum of node representations corresponding to $\hat{k}^y$ as a positive sample. Then, the gradient of the contrastive loss is passed to Emp-RFT.

6 Experiments

6.1 Dataset and Baselines

Dataset. Experiments were conducted on EmpatheticDialogues (Rashkin et al., 2019) which contains 24,850 multi-turn dialogues. For each dialogue, we can extract a certain number of instances corresponding to the number of turns within the dialogue. This totals to 47,611 instances, where 22,761 are multi-turn. In one turn of a dialogue, a speaker talks about one of 32 evenly distributed emotion labels and a situation related to the emotion label and a listener empathizes by responding to the speaker. Following the instructions of the dataset, we use 8:1:1 train/valid/test split.

Baselines. We compared Emp-RFT to the following five baseline models: (1) **MoEL** (Lin et al., 2019) is a transformer-based generative model, which has decoders for each emotion and integrates outputs of the decoders according to predicted emotion distribution. (2) **EmpDG** (Li et al., 2020) uses emotional words and consists of an adversarial framework including a generator and discriminators which reflect the user feedback. (3) **MIME** (Majumder et al., 2020) is also a transformer-based generative model which mimics user emotion based on emotion grouping and uses stochastic sampling for varied responses. (4) **MIME+Focused S1** and (5) **Blender+Focused S1** (Kim et al., 2021) attach RSA Framework to MIME and Blender (Roller et al., 2021). Blender is a pretrained model with 90M parameters size, using an immense number of dialogues. It is finetuned on EmpatheticDialogues. Using distractors and Bayes' Rules, RSA Framework makes the models focus on certain parts of the post, such as emotion cause words when generating responses in the single-turn dialogues [5]. Implementation details about Emp-RFT and baselines are covered in Appendix A.1.

6.2 Evaluation Metrics

Automatic Evaluation. We evaluated the models, using the following three metrics: (1) Perplexity (**PPL**) (Vinyals and Le, 2015) measures how highly likely tokens are generated, which evaluates the overall quality of the model. (2) Distinct-n (**Dist-n**) (Li et al., 2016) measures how diverse the generated response is via the unique words within its n-gram. (3). We use BERTscore (**F_{BERT}**) (Zhang et al., 2019) which measures token-level semantic similarities between the generated response and the gold response based on embeddings from BERT (Devlin et al., 2019).

Human Ratings. Human evaluations for the dialogues models are essential because of insufficient reliability on automatic metrics. We randomly sampled 100 test dialogues and asked 3 human workers to score models' generated responses on 1 to 5 point scale, following the four metrics (Rashkin et al., 2019): (1) **Empathy** measures whether the generated response understands the speaker's emotion and situation. (2) **Relevance** measures whether the generated response is coherent to the context. (3) **Fluency** measures whether the generated response is grammatically correct and readable. (4) Since we conclude that models generating generic responses are not empathizing to the speaker, we

[5] To make the models work in the multi-turn dialogues, the models are converted to take several utterances and to focus on emotion cause words of the last utterance.

Method	Automatic Evaluation				Human Evaluation			
	PPL	Dist-1	Dist-2	F_{BERT}	Empathy	Relevance	Fluency	Diversity
MoEL	38.04	0.44	2.10	0.11	3.25	3.73	3.49	2.85
EmpDG	37.29	0.46	2.02	0.14	3.30	3.76	3.57	3.11
MIME	37.09	0.47	1.91	0.13	3.23	3.78	3.53	2.83
MIME+Focused S1	36.43	0.52	2.21	0.15	3.34	3.84	3.65	3.15
Blender+Focused S1	**13.21***	3.11*	4.38*	0.31*	3.69*	4.11*	4.05*	3.78*
Emp-RFT	13.59*	**3.24***	**4.59***	**0.34***	**3.78***	**4.23***	**4.11***	**4.02***
w/o FTR	15.22	3.22*	4.49*	0.27	3.56	4.05	4.01	3.95*
w/o CP	13.89*	3.07	4.36	0.33*	3.74*	4.20*	4.05*	3.84*
w/o (CP+NEKD)	14.87	2.89	4.08	0.28	3.61	4.02	3.95	3.69
w/o (CP+NEKD+FCK)	15.45	2.75	3.86	0.23	3.51	3.89	3.83	3.50
MoEL	41.13	0.40	1.96	0.08	2.97	3.44	3.30	2.55
EmpDG	40.10	0.41	1.91	0.11	3.01	3.50	3.32	2.85
MIME	40.51	0.42	1.82	0.09	2.94	3.51	3.29	2.53
MIME+Focused S1	39.58	0.48	2.11	0.11	3.05	3.59	3.39	2.91
Blender+Focused S1	16.96	3.03*	4.19*	0.26*	3.43*	3.90*	3.88*	3.65*
Emp-RFT	**14.71***	**3.21***	**4.48***	**0.32***	**3.66***	**4.15***	**4.01***	**3.91***
w/o FTR	17.12	3.20*	4.40*	0.22	3.32	3.83	3.85	3.88*
w/o CP	15.12*	3.04*	4.28*	0.31*	3.62*	4.11*	3.96*	3.71*
w/o (CP+NEKD)	16.24	2.84	4.02	0.25*	3.50*	3.92*	3.84	3.61
w/o (CP+NEKD+FCK)	17.33	2.71	3.78	0.20	3.42	3.82	3.76	3.42

Table 2: Results of automatic evaluation and human ratings on all(top) and multi-turn(bottom) instances. * means superior results with p-value < 0.05 (sign test).

Emp-RFT vs.	Win (%)	Lose (%)	κ
MoEL	74.4/82.2	9.3/6.9	0.67/0.73
EmpDG	70.3/77.7	12.3/9.8	0.61/0.70
MIME	71.6/79.5	11.1/8.2	0.64/0.71
MIME+Focused S1	65.3/74.5	13.2/10.8	0.61/0.66
Blender+Focused S1	32.0/38.6	25.5/22.5	0.46/0.48

Table 3: Results of human A/B test. The results in front of and behind '/' are each on all instances and multi-turn instances. Fleiss' kappa (κ) denotes agreements among human workers, where $0.4 < \kappa < 0.6$ and $0.6 < \kappa < 0.8$ indicate moderate and substantial agreements, respectively.

use **Diversity** to measure whether the generated response is non-generic.

Human A/B Test. We further conducted a human A/B test which provides stronger intuitions and higher agreements than human ratings, because this is carried with 3 human workers selecting the better response when given two generated responses (Sabour et al., 2021).

6.3 Analysis of Response Generation

We abbreviate feature transition recognizer, contrastive PPLM, next emotion and keywords detection, and fusing context with keywords as FTR, CP, NEKD, and FCK, respectively.

Automatic Evaluation Results. The overall automatic evaluation results are shown in the left part of Table 2. Emp-RFT performed exceedingly on all metrics except for PPL, which was nearly the same as Blender+Focused S1. The improvements on other metrics indicated that our approach was effective for generating generally high quality and non-generic responses which were also semantically similar with the gold response. While the utilization of pretrained models yielded significant improvements compared to models only trained on EmpatheticDialogues, Emp-RFT showed even greater performance when compared to Blender+Focused S1 endowed with more significant number of dialogues. In addition, due to utilization of FTR, Emp-RFT obtained remarkable results even on multi-turn instances, whereas, other models suffered due to their means of utilizing features for the context at the coarse-grained level.

Human Evaluation Results. In the right part of Table 2, Emp-RFT acquired the highest scores on all metrics, which demonstrated that all components of Emp-RFT helped generate responses that are empathetic, coherent to the context, and non-generic. Also, utilizing pretrained models showed significant improvements, especially on Fluency and Diversity scores. In Table 3, the generated responses from Emp-RFT were more preferred, which indicated Emp-RFT consistently outperformed other methods in various experiments. When observing at the models' performance difference between multi-turn instances and all instances, only Emp-RFT continued to perform consistently, whereas other models showed significant perfor-

Method	Top-1 Acc	Top-5 Acc	macro-F1
CoMAE	41.12	80.09	39.61
Emp-RFT	**42.08**	**80.78**	**40.39**
Method	**TL-P**	**TL-R**	**TL-F1**
ConceptFlow	37.68	48.27	42.32
CG-nAR	44.62	61.94	51.87
Emp-RFT	**45.35**	**63.15**	**52.78**

Table 4: The results of next emotion(top) and keywords(bottom) detection. We used the metrics introduced in Table 1.

mance drops under multi-turn instances. From this, we concluded that Emp-RFT continuously understood the dialogue flow.

Ablation Study. To better understand effects of each component in Emp-RFT, we conducted the ablation study. We gradually ablated each component within the response generation strategy in a hierarchical manner. (1) **w/o FTR**: Feature transition recognizer was disabled, which resulted in considerable drops on all metrics, especially on PPL, F_{BERT}, Empathy, and Relevance scores on multi-turn instances, because Emp-RFT could not grasp the attention-needed features of utterance within multi-turn instances through FTR. (2) **w/o CP**: Contrastive PPLM was removed, which caused lower Dist-n and Diversity scores, because Emp-RFT could not actively use various $\hat{k}^y$ when generating responses through CP. (3) **w/o (CP+NEKD)**: Next emotion and keywords detection were disabled, which interfered with Emp-RFT's utilization of the next emotion and keyword. It droped not only Dist-n and Diversity scores but also other metrics. (4) **w/o (CP+NEKD+FCK)**: Fusing the context representation with keyword representations was disregarded. Since keywords were no longer emphasized for context modelling, such information could not get more attention when generating responses. It caused drops on all metrics, particularly on F_{BERT}, Dist-n, and Diversity.

6.4 Analysis of Next Emotion and Keywords

We report the results in terms of NEKD in Table 4. Since all baselines have not conducted NEKD, we trained models showing promising results such as **CoMAE** (Zheng et al., 2021), **ConceptFlow** (Zhang et al., 2020) and **CG-nAR** (Zou et al., 2021) with EmpatheticDialogues. (More details are covered in Appendix A.2). Then, we compared Emp-RFT to those models. Emp-RFT outperformed other models on all metrics, which proved Emp-RFT figured out which emotion and keywords were proper for generation.

Emotion Label: Furious
Annotated Emotion: Annoyed → Apprehensive → Confident → Hopeful
u^1: My roommate eats my food sometimes. This makes me so angry!
u^2: You should get a mini fridge and put it in your room, with a lock on it.
u^3: I think that's a great idea. I know where to get those fridges for cheap.
Gold: Yea man go for it, don't procrastinate.

MoEL: I am sorry you have to hear that. I hope it works out for you.
EmpDG: I agree with you. I think it isn't worth before you get it back.
MIME: I am sorry to hear that. I hope you don't have to deal with that.
MIME+Focused S1: I agree. I have a friend who is not to be a parent.
Blender+Focused S1: Roommates can be so annoying.

$\hat{e}^y$: Trusting, $\hat{k}^y$: procrastinate, safety, profit
Emp-RFT: Don't procrastinate. It makes your foods safety.
w/o FTR: I agree. I have a safety but pricy fridge.

Emotion Label: Proud
Annotated Emotion: Trusting → Proud → Grateful→ Sentimental → Disappointed → Sentimental
u^1: Hi, My girlfriend was recently let go of her job, but she put so much work in to finding a new job and got one with better pay!
u^2: That's great! It was meant to happen then.
u^3: Yes I agree, she was so happy when she found out the job was hers.
u^4: I would have been too. I've never been fired. I always quit.
u^5: She wasn't really fired, her position was no more. Schools don't have much room left for Home Ec teachers. Sad.
Gold: I remember taking Home Ec. I loved that class.

MoEL: I am sorry to hear that. I hope you can find a new job.
EmpDG: I am sorry you had that. I am not a fan of the past few years ago. I am not sure she had a close relationship.
MIME: I am sorry to hear that. I hope you find a new job.
MIME+Focused S1: I am sorry to hear that. I hope you find a new job.
Blender+Focused S1: She was paid for her hard work.

$\hat{e}^y$: Sentimental, $\hat{k}^y$: Ec, class, jobs, solution, remember
Emp-RFT: I remember the Ec class. It's a relief for her to find the job.
w/o FTR: I'm glad she found the solution to the job problem.

Emotion Label: Caring
Annotated Emotion: Annoyed → Caring → Sad→ Caring
u^1: My poor neighbor is suffering a lot without her husband.
u^2: I suffer a lot too when my wife is gone. What happened to him?
u^3: He passed away from cancer.
Gold: Ah, the evil cancer. Took my grandmother as well. I am sure he's off in a better place now.

$\hat{e}^y$: Sad, $\hat{k}^y$: condolences, cancer, grandmother, evil, lost
Emp-RFT: My condolences. I lost my grandmother because of the cancer.
w/o CP: I'm sorry to hear that. It's so hard to lost someone.
w/o (CP+NEKD): Oh no. I'm scary to get the cancer.
w/o (CP+NEKD+FCK): Oh no. I'm sorry to hear that.

Table 5: Model generations. The words marked in red and blue are keywords of the speaker utterance and listener utterance, respectively.

6.5 Case Study

The cases from the models are shown in Table 5. In the first case, MoEL and MIME expressed regret, which was emotionally inappropriate to the context. All baselines except for MoEL failed to grasp the proper features within the context, and therefore generated incoherent responses. Especially, Blender+Focused S1 ignored the features of u^3. Since Emp-RFT understood the dialogue flow, it became attentive to not only the features of u^3 but also those of u^1, u^2, mentioning ('procrastinate', 'foods', 'safety'), which led to empathy and coherence. In the second case, all baselines couldn't understand the longer context, which resulted in im-

proper empathy. Also, Blender+Focused S1 disregarded the features of u^5, and therefore overlooked the speaker's sadness. Emp-RFT fully comprehended why the speaker's happiness changed to the sadness. In both cases, without FTR, the responses of Emp-RFT were non-empathetic and incoherent because of dismissing appropriate features. In the third case, we report the case in terms of the response generation strategy. Without CP, NEKD, and FCK, Emp-RFT produced a generic response. With the utilization of FCK, Emp-RFT perceived the word 'cancer' in u^3 but expressed excessive emotion by mentioning 'scary'. When Emp-RFT additionally conducted NEKD, Emp-RFT generated emotionally appropriate responses by mentioning 'sorry' and 'hard', and utilized the keyword 'lost'. Lastly, with CP, Emp-RFT generated a diverse response, actively using $\hat{k}^y$.

7 Conclusion

We proposed a novel approach that recognizes feature transitions between utterances, which led to understanding the dialogue flow and grasping the features of utterance that needs attention. Also, to make our model focus on emotion and keywords related to appropriate features, we introduced a response generation strategy including fusing context with keywords, next emotion and keywords detection, and contrastive PPLM. Experimental results showed that our model outperformed baselines, and especially, achieved significant improvements on multi-turn instances, which proved our approach was effective for empathetic, coherent, and non-generic response generation.

8 Ethical Considerations

We expect that our proposed approach does not suffer from ethical problems. The dataset we use in our work is EmpatheticDialogues which is English-based. The dataset is constructed by crowdsourcing with Amazon Mechanical Turk, which protects private user information (Rashkin et al., 2019). In addition, the dialogue dataset is anticipated not to have responses which include discrimination, abuse, bias, etc, because the robust collection procedure of EmpatheticDialogues ensures the quality of the dataset. Thus, we expect that models trained using the dataset, do not generate inappropriate responses which harm the users. However, we inform that our model utilizes a pretrained language model, which may produce inappropriate responses. Lastly, we anticipate our model make potential users be interested and consoled by generating empathetic responses.

Acknowledgements

We thank all anonymous reviewers for their meaningful comments, and Hyeongjun Yang, Chanhee Lee and Sunwoo Kang of Yonsei University for their discussion and feedback about our work. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP; Ministry of Science, ICT & Future Planning) (No. NRF-2022R1A2B5B01001835). Also, this work was partly supported by the Institute of Information and Communications Technology Planning and Evaluation(IITP) grant funded by the Korean government(MSIT) (No. 2020-0-01361-003, Artificial Intelligence Graduate School Program (Yonsei University)). Kyong-Ho Lee is the corresponding author.

References

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Nancy Eisenberg and Janet Strayer. 1987. Critical issues in the study of empathy.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

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

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12911–12919.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

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

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation.

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)*, pages 121–132.

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

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312.

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

Lisong Qiu, Yingwai Shiu, Pingping Lin, Ruihua Song, Yue Liu, Dongyan Zhao, and Rui Yan. 2020. What if bots feel moods? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1161–1170.

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.

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

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. Cem: Commonsense-aware empathetic response generation. *arXiv preprint arXiv:2109.05739*.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. 2021. Constructing emotional consensus

and utilizing unpaired data for empathetic dialogue generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3124–3134.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Shuohang Wang and Jing Jiang. 2017. A compare-aggregate model for matching text sequences.(2017). In *ICLR 2017: International Conference on Learning Representations, Toulon, France, April 24-26: Proceedings*, pages 1–15.

Haolan Zhan, Hainan Zhang, Hongshen Chen, Zhuoye Ding, Yongjun Bao, and Yanyan Lan. 2021. Augmenting knowledge-grounded conversations with sequential knowledge transition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5630.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL 2021*.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2226.

A Implementation Details

A.1 Empathetic Response Generation Models

We use the official codes of all baselines, and follow the implementations (MoEL [6], EmpDG [7], MIME [8], MIME+Focused S1 and Blender Focused S1 [9]). Our model is implemented by Pytorch [10], and based on two encoders of BART-base and a decoder of BART-base [11]. Hidden size d is 768 and the number of emotion classes n_{emo} is 32. MH and the number of layers of graph attention network are each 4. Using Adam optimization (Kingma and Ba, 2015), our model is trained on single RTX 3090 GPU with a batch size of 4. We apply early-stopping and select a model showing the best performance through perplexity on the valid set. For contrastive PPLM, we utilize the official code of PPLM [12]. We set a temperature parameter τ and batch size to 0.5 and 64, respectively. Through represenations derived from the last token of BART decoder whose parameters are frozen, we can obtain each response representation r_a and each keyword set representation ks_a, where the keyword set corresponds to the response. Thus, ks_a becomes a positive sample for r_a, and keyword set representations for other responses in the same batch become negative samples.

A.2 Next Emotion and Keywords Detection

We utilize the repositories and follow implemetation details of CoMAE [13], ConceptFlow [14], and CG-nAR [15]. We train three models, using EmpatheticDialogues instead of originally used datasets.

[6] https://github.com/HLTCHKUST/MoEL
[7] https://github.com/qtli/EmpDG
[8] https://github.com/declare-lab/MIME
[9] https://github.com/skywalker023/focused-empathy
[10] https://pytorch.org/
[11] https://huggingface.co/docs/transformers/model_doc/bart
[12] https://github.com/uber-research/PPLM
[13] https://github.com/chujiezheng/CoMAE
[14] https://github.com/thunlp/ConceptFlow
[15] https://github.com/RowitZou/CG-nAR