Improving Empathetic Dialogue Generation by Dynamically Infusing Commonsense Knowledge

Hua Cai[*†] **Xuli Shen**[*] **Qing Xu** **Weilin Shen** **Xiaomei Wang**
Weifeng Ge **Xiaoqing Zheng** **Xiangyang Xue**
UniDT Technology, Shanghai, China
School of Computer Science, Fudan University, Shanghai, China

Abstract

In empathetic conversations, individuals express their empathy towards others. Previous work has mainly focused on generating empathetic responses by utilizing the speaker's emotion. Besides, external commonsense knowledge has been applied to enhance the system's understandings of the speaker's situation. However, given an event, commonsense knowledge base contains various relations, potentially leading to confusion for the dialogue system. Consequently, inconsistencies arise among the emotion, generated response and speaker's contextual information. To this end, we propose a novel approach for empathetic response generation, which incorporates an adaptive module for commonsense knowledge selection to ensure consistency between the generated empathetic responses and the speaker's situation. This selected knowledge is used to refine the commonsense cognition and empathy expression for generated responses. Experimental results show that our approach significantly outperforms baseline models in both automatic and human evaluations, exhibiting the generation of more coherent and empathetic responses. Moreover, case studies highlight the interpretability of knowledge selection in the responses and the effectiveness of adaptive module in our model. Code: https://github.com/Hanscal/DCKS.

1 Introduction

Empathy is a desirable human ability in our daily conversations. It is known as a complex multi-dimensional construct encompassing social, cognitive, and emotional processes, which enables us to experience the emotion of others through various emotional stimuli and to understand the implicit mental states of others (Davis, 1983; Zheng et al., 2021). Previous research (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020; Li et al., 2021b) has been conducted on dialogue systems to enhance its empathy ability in open-domain. In order to generate empathetic responses, one line of growing interests is incorporating commonsense knowledge into conversation modeling (Ghosal et al., 2020; Zhou et al., 2021; Sabour et al., 2021).

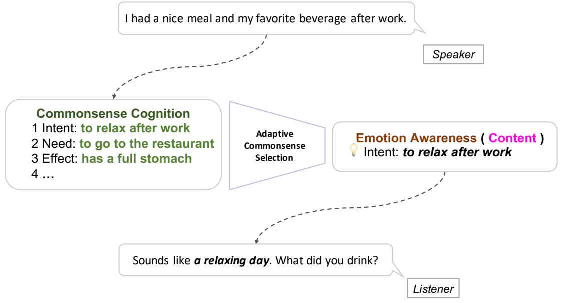
I had a nice meal and my favorite beverage after work.
Speaker
Commonsense Cognition
1 Intent: to relax after work
2 Need: to go to the restaurant
3 Effect: has a full stomach
4 ...
Adaptive Commonsense Selection
Emotion Awareness (Content)
Intent: to relax after work
Sounds like a relaxing day. What did you drink?
Listener

Figure 1: The framework of our proposed empathetic dialogue generation. The listener acknowledges speaker's feeling with the adaptive commonsense selection and respond with respect to the emotion status of speaker.

Yet, understanding speaker's emotion and showing the contextually appropriate comprehension of her/his situation are still challenges in empathetic conversations. When interacting with a dialogue system, the speakers are not expected to explicitly share all the information about their situation and how they may feel. As humans, we use our commonsense knowledge to make connections between what is explicitly mentioned and what is implied. Hence, to address above issues, some prior works (Zhou et al., 2018b; Wu et al., 2020) implement external knowledge to identify the speaker's situation, to acknowledge the speaker's status and to bring diversity for generated response.

However, straightforward knowledge merging method confuses the system and the response consistency would be deteriorated. This is demonstrated in Figure 1, where the irrelevant knowledge (*Need*) may potentially form empathetic responses, which conflicts with the information about speaker's emotion (*content*). Accordingly, the speaker displays the satisfaction of her/his expe-

[*] These authors contributed equally to this work.
[†] Corresponding author: Hua Cai (hua.cai@unidt.com)

Findings of the Association for Computational Linguistics: ACL 2023, pages 7858–7873
July 9-14, 2023 ©2023 Association for Computational Linguistics

rience, which provides potential informative cognitions based on one unified commonsense. We can assume that if the most appropriate commonsense cognition (*Intent*) is selected with respect to emotion status, the generated response shows better consistency and empathy. Therefore, we believe dialogue systems with rectified knowledge, which aims at unifying the contextual emotion, lead to more consistent and empathetic responses.

In this paper, we address the task of empathetic dialogue generation by dynamically infusing commonsense knowledge. Such additional commonsense knowledge is used to improve the cognitive understanding about the speaker's situation and feelings, thus enhance the empathy expression in the generated responses. Meanwhile, the dynamical selection stage avoids the confusion of knowledge in dialogue system and enhance the response consistency with context history. In general, our main contributions are summarized as follows:

- We introduce a novel approach that incorporates the inferred commonsense knowledge to enhance empathetic response generation.
- We propose an effective knowledge selecting paradigm that could dynamically select the commonsense knowledge, which is most relevant to speaker's cognitive empathy. To the best of our knowledge, it is the first work to study commonsense knowledge dynamical selection for empathetic dialogue generation.
- Experiments show that with incorporating the selected commonsense, our model is able to generate more empathetic and interpretable responses compared with the previous methods.

2 Related Works

2.1 Empathetic Dialogue Generation

In recent years, research on implementing empathy in open domain dialogue systems and generating empathetic responses has gained considerable attention. Rashkin et al. (2019) consider a richer and evenly distributed set of emotions and release a dataset EmpatheticDialogues, where a listener responds to a speaker who is under an emotional situation in an empathetic way. Ghosal et al. (2020) demonstrate that detecting the speaker's emotion is an essential part of generating empathetic responses. Prior studies on emotion-related conversational systems mainly focused on rule-based systems, which heavily rely on hand-craft features (Zhou and Wang, 2018; Zhou et al., 2018a). Recently, many neural emotional dialogue generation approaches have been explored to control the emotional expression in the target response (Lin et al., 2019; Majumder et al., 2020). However, Li et al. (2021a) reveal that conventional empathetic conversation systems face an emotional inconsistency problem as they strive to produce emotionally rich responses based on predefined user-input emotions.

2.2 Connecting Knowledge and Dialogue

Leveraging knowledge from commonsense knowledge base has been demonstrated for gaining a better understanding of the implied emotions within the context (Tu et al., 2022; Lee et al., 2022). ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019) are commonsense knowledge bases. ConceptNet consists of 36 relations focusing mostly on taxonomic, lexical and physical commonsense knowledge. Distinguished from ConceptNet, ATOMIC consists 9 relations that cover social commonsense knowledge including event-centered causes and effects as well as person-related mental states. Both Zhou et al. (2018b) and Zhang et al. (2019) introduce knowledge triplets from ConceptNet into open-domain response generation. Recently, Li et al. (2022) and Zhong et al. (2021) exploit ConceptNet to enhance emotion reasoning for response generation. Ghosal et al. (2020) utilizes ATOMIC in emotional dialogue modeling for emotion identification. Sabour et al. (2021) leverages commonsense from ATOMIC to improve the understanding of speaker's situations and feelings.

Therefore, enabling dialogue systems to leverage commonsense and driving implications from the speaker's explicit statements are highly beneficial for more empathetic responses. In this work, we focus on the task of empathetic dialogue generation on EmpatheticDialogues dataset, and pay attention to addressing social related commonsense knowledge from ATOMIC. For each event, we use the social relations in ATOMIC to infer the commonsense knowledge about the person involved in the event. We adopt COMET (Bosselut et al., 2019) to generate commonsense sentences for the given events. This model is pre-trained on triplets from ATOMIC and then fine tuned on ATOMIC$^{20}_{20}$ (Hwang et al., 2021), so that is more suitable for inferring knowledge regarding unseen events in the original ATOMIC daily basis dataset.

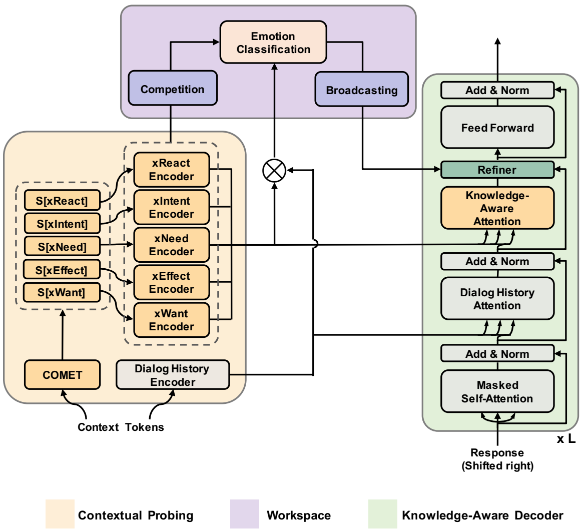

Figure 2: The architecture of our framework. It consists of three modules: (1) Contextual Probing enhances dialogue comprehension by commonsense knowledge; (2) Workspace adaptively modifies the cognition of speaker's status; (3) Knowledge-Aware decoder generates empathetic responses.

3 Methodology

Our proposed model is built upon the Transformer-based pre-trained language model to generate listener's utterance. Each conversation process of the model is mainly divided into three stages: contextual probing, contextual unification workspace and knowledge-aware decoder. The overview of our model is illustrated in Figure 2.

3.1 Task Formulation

The task requires a dialogue model to play the role of the listener and generate empathetic responses. Formally, let $U = [u_1,\ u_2,\ ...,\ u_{n-1}]$ denote a dialogue history of $n-1$ utterances, where $u_i = [w_1^i,\ w_2^i,\ ...,\ w_{M_i}^i]$ is the i-th utterance that consists of M_i words. Let $K = \{k_i\}$ denote the commonsense knowledge generated from COMET, where k_i is the empathetic commonsense inference knowledge. Our goal is to generate a response Y using historical utterance U and commonsense knowledge K as input. A dialogue history encoder to encode U, a knowledge encoder to encoder K, and a decoder to incorporate dialog history, dynamically select knowledge and generate response.

3.2 Contextual Probing

To obtain semantic representations of the dialog history and the knowledge from ATOMIC, we divide the context probing part into context encoding and knowledge acquisition.

3.2.1 Context Encoding

We concatenate the utterances in the dialogue history and prepend a special token $[CLS]$ to obtain the dialogue historical context input $U = [CLS] \oplus u_1 \oplus u_2 \oplus ... \oplus u_{n-1}$, where $\oplus$ is the concatenation operation. Then, we use the final hidden representation of $[CLS]$ as the representation of the whole sequence.

We use BART encoder part to acquire the contextual representation. The sequence U is fed into the encoder, and the hidden state of the encoder token:

$$\mathbf{z}_{\text{ctx}} = \mathbf{Enc}_{\text{ctx}}(U), \tag{1}$$

where $\mathbf{z}_{\text{ctx}} \in \mathbb{R}^{L \times d}$, L is the length of the sequence, and d is the hidden size of the context encoder.

3.2.2 Knowledge Acquisition

In ATOMIC, six relations could be inferred for the person X involved in the event: the effect of the event on X ($xEffect$), X's reaction to the event ($xReact$), X's intent before the event ($xIntent$), what X need in order for the event to happen ($xNeed$), what X would want after the event($xWant$), and an inferred attribute of X's characteristics ($xAttr$). Since predicting a person's attributes involves judging the other person, which is not included in the empathetic process, we ignore $xAttr$ in our approach and use the remaining five relations.

For input sequence U, we respectively append five special relation tokens ([xReact], [xWant], [xNeed], [xIntent], [xEffect]) to the last utterance in the dialogue history and then use COMET to generate k commonsense inferences $S^r = [cs_1^r,\ cs_2^r,\ ...\ ,\ cs_k^r]$ per relation r, where $r \in \{xReact,\ xWant,\ xNeed,\ xIntent,\ xEffect\}$.

For each relation, we concatenate the generated commonsense inferences to obtain its commonsense sequence $CS_r = cs_1^r \oplus cs_2^r \oplus ... \oplus cs_k^r$, which demonstrates the knowledge regarding the speaker's dialogue state (i.e. emotion and situation). Accordingly, similar to the previous section, we prepend $[CLS]$ to the sequences denoted as $\mathbf{E}_{CS_r}$, which then are fed to five separate commonsense knowledge encoders, as shown in the contextual probing part of Figure 2:

$$\mathbf{Z}_r = \mathbf{Enc}_{Kno}(\mathbf{E}_{CS_r}), \tag{2}$$

where $\mathbf{Z}_r \in \mathbb{R}^{l_r \times d}$, l_r is the lengths of the commonsense inference sequences.

Then, we utilize the hidden vector of $[CLS]$ as the representation for each relation, and through average operation we obtain the fused representation $\mathbf{z}_r = Average(\mathbf{Z}_r[0]) \in \mathbb{R}^d$ for all relations.

3.3 Contextual Unification Workspace

To better leverage the hidden representation from knowledge acquisition and context encoding, we apply the workspace module for unifying contextual information according to emotion label. The workspace consists of two parts: emotion classification for identifying speaker's status, and adaptive knowledge selection for excluding irrelevant knowledge representation.

3.3.1 Emotion Classification

In contrast to concatenating the representations at a sequence level, we use point-wise addition to fuse the additional knowledge in the sequence, i.e., the fusing of knowledge and the context representation:

$$\mathbf{z}_f = \mathbf{z}_r + \mathbf{z}_{\text{ctx}}. \tag{3}$$

In order to acquire a more accurate prediction of the speaker's emotion, given that we are provided with an emotion label e for each conversation, we use the infused representation of knowledge and context representation to perform emotion classification. We also pass $\mathbf{z}_f$ through a linear layer g_θ, followed by a softmax operation to produce the emotion category distribution $P_{\text{emo}} \in \mathbb{R}^q$, where q is the number of available emotion categories:

$$P_{\text{emo}} = \text{Softmax}(g_\theta(\mathbf{z}_f)), \tag{4}$$

where $\theta \in \mathbb{R}^{d \times q}$ is the weight vector for the linear layer. During training, we optimize these weights by minimizing the Cross-Entropy (CE) loss between the emotion category distribution P_{emo} and the ground truth label e:

$$\mathcal{L}_{\text{emo}} = -\log(P_{\text{emo}}(e)). \tag{5}$$

3.3.2 Adaptive Knowledge Selection

We present a knowledge selection method that the decoder can adaptively choose the commonsense representations based on the emotion classification results. Given the set of knowledge representation $\mathbf{Z} = \{\mathbf{Z}_r[0]\}$, the goal is to choose the most appropriate knowledge relations that satisfy the consistency with the context representation vector $\mathbf{z}_{\text{ctx}}$. By this selection paradigm, the irrelevant relations, which would potentially confused the generated response, will be eliminated, so as to boost the performance of dialogue system.

Inspired by Global Workspace Theory in cognitive science (Blum and Blum, 2022; Baars, 1993) , the process of contextual coordination is realized by eliminating irrelevant cognition. We therefore implement the label of emotion as the coordination of context and the $\mathcal{L}_{\text{emo}}(g_\theta(\mathbf{z}), g_\theta(\mathbf{z}_{\text{ctx}}))$ from the supervised evaluation to eliminate irrelevant cognition. The knowledge selection mechanism is divided into two stages, *competition* and *broadcasting*:

- During the *competition* stage, we recursively exclude the irrelevant information of knowledge representation based on the emotion status. Specifically, at iteration m, we choose the $\max_{\mathbf{z} \in \mathbf{Z}}\{\mathcal{L}_{\text{emo}}(g_\theta(\mathbf{z}), g_\theta(\mathbf{z}_{\text{ctx}}))\}$ as the most irrelevant knowledge representation. In order to model the influence of knowledge exclusion, we leverage nonlinear regression method (Xu and Xuan, 2019; Shen et al., 2022) to calculate the dynamics $\mathbf{G} = \nabla_{\boldsymbol{\theta}} \mathbf{f} \in \mathbb{R}^{d \times q}$ of the aforementioned max loss. Please refer to the Appendix for the technical details. After the last iteration, the remaining knowledge representation, as the winner of competition, is applied for acknowledging the unified speaker's emotion status.

- In the *broadcasting* stage, the winner of competition stage will be applied for unifying the combined representation in decoder. Specifically, we realize this stage by adding the dynamics of the selecting process to rectify the knowledge representation. Thus, the generated response will less affected by the unrelated information from knowledge encoder in contextual probing module.

We provide Algorithm 1 in Appendix to show the exclusion method. Figure 3 displays how the workspace process refine the knowledge representation.

3.4 Knowledge-Aware Decoder

Generally, not all knowledge contributes to the generation of the response, so the model should have the ability to select knowledge. Instead of performing knowledge selection in the encoding phase, we leave it to the decoding phase. As shown in the right part of Figure 2, a knowledge-aware

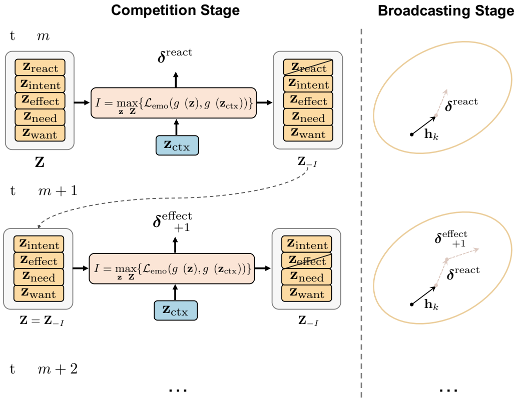
Competition Stage
Broadcasting Stage
t m
$\mathbf{z}_{\text{react}}$
$\mathbf{z}_{\text{intent}}$
$\mathbf{z}_{\text{effect}}$
$\mathbf{z}_{\text{need}}$
$\mathbf{z}_{\text{want}}$
$\mathbf{Z}$
$\boldsymbol{\delta}^{\text{react}}$
$I = \max_{\mathbf{z}\ \mathbf{Z}}\{\mathcal{L}_{\text{emo}}(g\ (\mathbf{z}), g\ (\mathbf{z}_{\text{ctx}}))\}$
$\mathbf{z}_{\text{ctx}}$
$\mathbf{Z}_{-I}$
$\mathbf{h}_k$
t $m+1$
$\boldsymbol{\delta}^{\text{effect}}_{+1}$
$\mathbf{Z} = \mathbf{Z}_{-I}$
t $m+2$
…

Figure 3: The illustration of the workspace mechanism. During competition stage at each iteration, the most irrelevant knowledge, for example the 'react', is deleted from the set of knowledge representation, demonstrated by $\mathbf{Z}_{-I}$. The dynamics of the deletion is $\boldsymbol{\delta}$. During broadcasting stage, the knowledge-aware representation $\mathbf{h}_k$ is refined by the dynamic $\boldsymbol{\delta}$.

cross attention block is introduced to select knowledge dynamically. Feed the selected knowledge to the context-knowledge refiner, which assists in response generation. The fused knowledge is taken as the input of this block, and then the output of this block is refined to exploit the knowledge contributions.

3.4.1 Knowledge Refiner

In order to refine the context and knowledge contributions in each layer, we replace the residual addition to a refine gate after the knowledge-aware attention block. Denote $\mathbf{h}_k$ as output of knowledge-aware attention block and $\mathbf{h}_c$ as the residual from the previous block, the output of refiner can be expressed by:

$$R_f(\widetilde{\mathbf{h}}_k, \mathbf{h}_c) = \alpha \cdot \mathbf{LN}(\widetilde{\mathbf{h}}_k) + (1-\alpha) \cdot \mathbf{h}_c \quad (6)$$

$$\widetilde{\mathbf{h}}_k = \mathbf{h}_k + \boldsymbol{\delta}_m \quad (7)$$

$$\alpha = \sigma(\mathbf{w} \cdot [\widetilde{\mathbf{h}}_k; \mathbf{h}_c]) \quad (8)$$

Where $\mathbf{LN}$ is a linear layer, $\widetilde{\mathbf{h}}_k$ is the rectified knowledge representation, $\mathbf{w} \in \mathbb{R}^{2d}$ is a learnable parameter and σ denotes sigmoid function.

3.4.2 Response Generation

Lastly, the target response $Y = [y_1,\ y_2,\ \ldots,\ y_T]$ with length T, which is generated by the decoder token by token by using the embeddings of the tokens that have been generated and the commonsense-refined contextual representation $R_f(\widetilde{\mathbf{h}}_k, \mathbf{h}_c)$, which has fused the information from both the context and the commonsense inferences. We adopt the standard negative log-likelihood (NLL) loss on the target response Y:

$$\mathcal{L}_{\text{nll}} = -\sum_{t=1}^{T} \log(y|(\mathbf{U}, \mathbf{K}), y_{<t}). \quad (9)$$

3.5 Training Objectives

All the parameters for our proposed model are trained and optimized based on the weighted sum of the two mentioned losses:

$$\mathcal{L} = \mathcal{L}_{\text{nll}} + \gamma\mathcal{L}_{\text{emo}}, \quad (10)$$

where γ is hyper-parameter that we use to control the influence of the these losses. In our experiments, we set $\gamma = 1$.

4 Experiments

4.1 Datasets

We conduct our experiments on the EmpatheticDialogues, a large-scale multi-turn dataset containing 25k empathetic conversations between crowd sourcing workers. The dataset also provides an emotion label for each conversation from the total 32 available emotions.

4.2 Baselines

We select the following baseline models for comparison on EmpatheticDialogues: (1) **Transformer** (Vaswani et al., 2017): An original Transformer, which is trained to optimize the NLL loss. (2) **Multi-TRS** (Rashkin et al., 2019): A variation of the Transformer for multitask that trained to jointly optimize an additional cross-entropy loss for emotion classification with the NLL loss. (3) **MoEL** (Lin et al., 2019): A Transformer-based model that uses 32 emotion-specific decoders to generate a response. Therefore, each decoder is optimized to respond appropriately for each emotion. (4) **MIME** (Majumder et al., 2020): Another Transformer-based model that mimics the context emotion to a varying degree considering its negative and positive emotions, and then generates empathetic response based on the blend of these two emotions. (5) **EmpDG** (Li et al., 2021a): A multi-resolution adversarial framework which applies an empathetic generator to produce empathetic responses and an interactive discriminator to ensure that the generated responses are consistent with the context and are also empathetic. (6) **CEM**

Models	PPL	B-1	B-2	B-3	B-4	R-1	R-2	Dist-1	Dist-2	Acc
Transformer	37.62	18.07	8.34	4.57	2.86	17.22	4.21	0.36	1.35	–
Multi-TRS	37.50	18.78	8.55	4.70	2.95	16.85	4.21	0.35	1.27	33.95
MoEL	36.60	18.07	8.30	4.37	2.65	18.24	4.81	0.59	2.64	31.74
MIME	37.24	18.60	8.39	4.54	2.81	17.08	4.05	0.47	1.66	30.96
EmpDG	37.43	19.96	9.11	4.74	2.80	18.02	4.43	0.46	1.99	31.65
CEM	36.33	16.12	7.29	4.06	2.03	15.77	4.50	0.62	2.39	36.84
Ours	16.08	**21.73**	**10.62**	**6.24**	**4.09**	**19.77**	**5.65**	**2.19**	**9.61**	**49.16**
w/o A∗	15.41	19.50	9.54	5.52	3.62	19.35	5.57	2.16	8.87	46.47
w/o Knowledge	**15.24**	20.11	9.86	5.72	3.73	19.72	5.82	2.08	8.59	44.87
w/o Context	15.62	20.45	9.98	5.78	3.74	19.88	5.78	1.82	7.41	46.34

Table 1: Results of automatic evaluation. A∗ represents the adaptive knowledge selection method in the workspace module.

(Sabour et al., 2021): An empathetic generation approach which leverages commonsense to draw more information about the speaker's situation and uses this additional information to further enhance the empathy expression in generated responses.

4.3 Implementation Details

We implement all the models using PyTorch and use the encoder and decoder from base version of BART in our work. We use Adam optimizer with initial learning rate 0.00005 in 5 epochs. The batch size is 16. The max sequence length in source and target is 256 and 64 respectively. We use the same 8:1:1 train/valid/test split as provided by Rashkin et al. (2019). In each experiment, we apply an early stop mechanism to prevent the model from over fitting, and then report the test results of the optimal model on the test set. All our training and test results were performed on 32GB Tesla V100 GPU.

4.4 Evaluation Metrics

4.4.1 Automatic Evaluation

We employ Perplexity (PPL), corpus-level BLEU (B-n), sentence-level ROUGE (R-n) and Distinct-n (Dist-n) as our main automatic metrics. Perplexity represents the model's confidence in its set of candidate responses, with higher confidence resulting in a lower PPL. This can be used to evaluate the general quality of the generated responses. Response with higher BLEU and ROUGE is closer to the ground-truth. Distinct-n measures the proportion of unique n-grams in the generated responses and is commonly used to evaluate generation diversity. In addition, since our proposed model and most baseline models perform emotion classification as part of their training process, we also report the prediction accuracy (Acc).

4.4.2 Human Evaluation

Following the methods in CEM, we conduct an aspect-based pairwise preference test. That is, for a given context, we pair our model's response with a response from the baselines and ask annotators to give each response a rating score from four aspects: 1) Coherence (**Coh.**): which response is more coherent in content and relevant to the context; 2) Empathy (**Emp.**): which response shows more understanding of the speaker's situation and presents a more appropriate emotion; 3) Informativeness (**Inf.**): which response conveys more information about the context. 4) Continuity (**Con.**): which response ignites the speaker's more desire to continue the conversation. Then, we randomly sample 100 response pairs and totally shuffle the response order in each sample. We assign crowd sourcing workers to annotate each pair on a scale of 1 to 5.

4.5 Evaluation Results

4.5.1 Automatic Evaluation Results

Table 1 reports the evaluation results on automatic metrics. Ours model achieves the lowest perplexity, which suggests the overall quality of our generated responses is higher than the baselines, approximately 56% lower than CEM. In addition, our model also considerably outperforms the baselines in terms of Dist-n, BLEU-n and ROUGE-n, which highlights the diversity of the responses and the relevance between generated response and speaker's situation. In terms of emotion classification, our model had a much higher accuracy compared to the baselines, nearly 34% higher than CEM, which suggests the adaptive selection of commonsense knowledge is pivotal for detecting the speaker's emotion.

Table 2 reports the evaluation results on low-resource training set, and we have the following observations: (1) In the full-data scenario, our model achieves start-of-the-art performance by infusing commonsense knowledge, which means that the importance of knowledge in dialogue generation. Besides, reducing the number of training samples has effect on model performance, but not that much, for that even the model using 1/4 data still has the approximate values in PPL, BLEU-n, ROUGE-n and Dist-n compared with the model using full data. (2) In the 1/8 training data scenario, our model achieves the comparable performance with baselines even though them leveraged all training data. (3) Responses generated by our model have higher Dist-n in low-resources scenarios, which means that our model can better obtain information from multiple knowledge and generate more diverse texts.

4.5.2 Ablation Studies

We conduct ablation studies to verify the effectiveness of each of the components in emotion classification and the generation performance. Specifically, we design three variants: *workspace*, *knowledge* and *context*. It is worth noting that since *workspace* depends on *knowledge* and *context*, when *knowledge* or *context* module is removed, *workspace* is removed by default:

1. w/o Adapter: the mechanism in workspace that used for adaptive commonsense knowledge selection is removed, and the emotion classification is based on none selected commonsense representation;

2. w/o Knowledge: the commonsense knowledge representation used for emotion classification is removed (Equation 6), and the hidden representation of the [CLS] token from the encoded context is used for emotion classification;

3. w/o Context: the context representation used for emotion classification is neglected (Equation 6), but keep the affective and cognitive commonsense knowledge representations;

The obtained results are shown in Table 1. We observe that reducing the workspace module results in lower classification accuracy as the same as BLEU-n and ROUGE-n. And removing the commonsense knowledge information also impacts the emotion classification accuracy. The above phenomena suggest that information about both the speaker's emotion and their situation are necessary for correctly identifying their feelings, and

Models	PPL	B-2	B-4	R-1	R-2	Acc
Ours	16.08	10.62	4.09	19.77	5.65	49.16
1/2 Data	16.57	10.00	3.58	19.56	5.35	40.00
1/4 Data	16.43	9.72	3.32	18.61	4.83	34.24
1/8 Data	18.51	9.33	3.29	18.61	4.82	33.80
1/16 Data	44.71	8.77	2.56	17.04	4.06	25.29
Zero Data	100+	3.99	0.86	10.20	1.09	2.60

Table 2: Evaluation results on low-resource training set of EmpatheticDialogues.

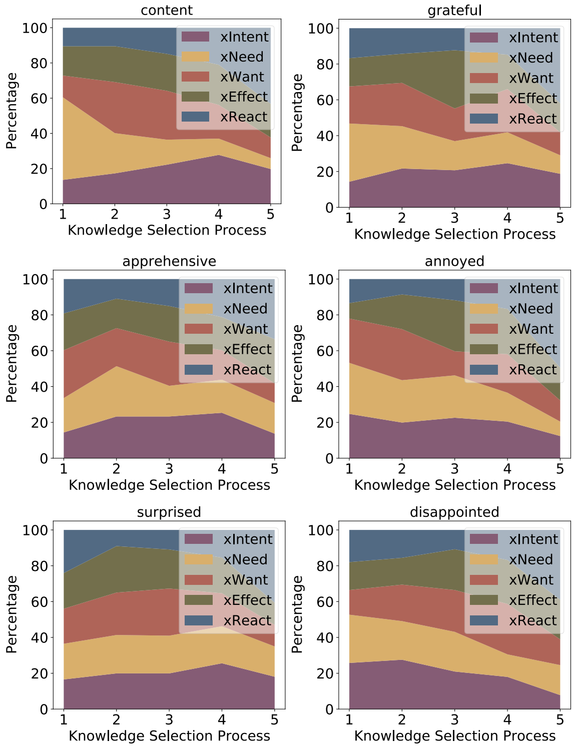

Figure 4: Stackplot of the knowledge selection process.

dynamical knowledge selection is leveraging the knowledge contribution to the cognition response. Removing those components leads to lower Dist-n scores but higher perplexity, which indicates the effectiveness of those components in generating more diverse responses.

4.5.3 Human Evaluation Results

Table 4 reports the evaluation results on human ratings. We observe that responses from our model are more contextually coherent than those from baselines. Besides, with the enhancement of commonsense knowledge, the response from our model are able to convey more specific and informative content. It is worth to note that, for the aspect of continuity, our model significantly outperforms all the baselines, which suggests that the generated responses may increase speaker's engagement, thus a more intimate emotional expression.

Context	I ended up ***staying at home*** for the holidays.
MoEL	Where did you go?
MIME	I am sorry to hear that. What happened?
EmpDG	Oh no! What happened?
CEM	Did you enjoy it?
Selection Process	xEffect → xReact → xWant → xNeed → xIntend
Selected Knowledge	xIntent: To be alone; ***To be home***; To relax; To rest
Emotion	***Lonely***
Ours	Oh, I am sorry to hear that. What did you do while staying at home ?
Gold	Why is that? Comfort?
Context	My friend borrowed quite a lot of money from me. I really ***do believe he will repay me***.
MoEL	That is so nice of him. Did you find a good friend?
MIME	That is a good thing to do.
EmpDG	That is a good friend.
CEM	That is nice of him.
Selection Process	xReact → xIntend → xEffect → xWant → xNeed
Selected Knowledge	xNeed: To ask for a loan; To get a loan; ***To ask him to repay***; To ask for money
Emotion	***Trusting***
Ours	I am sure he will repay you .
Gold	You do? That's good, friends can be terrible people to lend too.

Table 3: We report the case study of generated responses from EmpatheticDiaglogues. The responses with yellow background color demonstrate the awareness to the emotion and the selected knowledge.

Models	Coh.	Emp.	Inf.	Cont.
MoEL	3.57	3.26	3.11	3.09
MIME	3.61	3.30	3.09	3.13
EmpDG	3.42	3.10	2.94	2.89
CEM	3.90	3.49	3.08	3.19
Ours	**4.39**	**4.13**	**4.18**	**4.24**

Table 4: Results of human evaluation. We report the average scores of four aspects. Fleiss kappa of the results is 0.36, which constitutes a fair level of agreement.

4.6 Qualitative Studies

Case Study Table 3 shows the cases from EmpatheticDialogues, from which we can see that the response of our method outperforms the baselines. We analyze these cases with respect to the four factors evaluated by human. In aspect of *Coherence* and *Informativeness*, our response is more coherent in content and consistent to the context information. For instance, in case one, by the awareness of selected knowledge 'To be home', our method mentions this phrase in response so that the response better acknowledges speaker's intention. However, other methods fail to generate consistent response. It can be observed that MoEL and CEM dismiss the implication that the speaker is alone at home. The workspace module improves *Empathy* and *Continuity* by selecting the most influential commonsense with respect to the context. In both cases, the selected knowledge corresponds to the speaker's situation, which produces a more meaningful response by showing careness for speakers.

Efficacy of Knowledge Selection Selection process illustrates that the most irrelevant knowledge is selected and eliminated at each iteration. By combining dynamics from the selection process in refiner, the generated sentence gradually focuses on speaker's emotion status, so that our method provides more interpretable knowledge selection process for the dialogue system. Figure 4 provides characteristic of knowledge selection process. It indicates that workspace module tends to select inferred knowledge from the relation xReact. Since xReact reflects speaker's reaction to context, our adaptive selection method potentially provides the consistency between context and knowledge.

5 Conclusions

In this paper, we improve empathetic dialogue generation by infusing dynamical commonsense knowledge to promote the understanding of the speaker's situation and feelings, which leads to more consistent and empathetic responses. The automatic and human evaluation demonstrate that the effectiveness of our approach in high-quality empathetic response generation.

Limitations

One limitation in this work is the metrics employed in the automatic evaluation. The metrics mainly focus on the quality of generated response and the accuracy of emotion recognition, while automatic evaluation lacks a comprehensive method to evaluate empathy. Another limitation comes from the utilization of the dataset designed for open-domain dialogue system, so that the generated response from the proposed framework is not task-oriented. In the future, we will build empathetic dialogue generation datasets with diverse and task-oriented response, and develop metrics to evaluate the understanding of the speaker's situation.

Ethics Statement

The human evaluation is conducted by the employed workers, who does not involve privacy issues. We use public datasets to conduct our experiments. Existing packages involved in this work are displayed in the appendix.

Acknowledgment

This work was supported by UniDT's Cognitive Computing and Few Shot Learning Project.

References

Bernard J Baars. 1993. *A cognitive theory of consciousness*. Cambridge University Press.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Lenore Blum and Manuel Blum. 2022. A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. *Proceedings of the National Academy of Sciences*, 119(21):e2115934119.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

M. H. Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Jing Yang Lee, Kong Aik Lee, and Woon Seng Gan. 2022. Improving contextual coherence in variational personalized and empathetic dialogue agents. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7052–7056. IEEE.

Qintong Li, Hongshen Chen, Zhaochun Ren1, Pengjie Ren, Zhaopeng, and Zhumin Chen. 2021a. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.

Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. 2022. Towards empathetic dialogue generation over multi-type knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2021b. Knowledge Bridging for Empathetic Dialogue Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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 (EMNLP)*.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2021. Exemplars-guided Empathetic Response Generation Controlled by the Elements of Human Communication. In *CIKM*.

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)*.

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*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: Commonsense-aware Empathetic Response Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial intelligence*, volume 33, pages 3027–3035.

Xuli Shen, Xiaomei Wang, Qing Xu, Weifeng Ge, and Xiangyang Xue. 2022. Towards scalable and fast distributionally robust optimization for data-driven deep learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 448–457.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI Conference on Artificial Intelligence*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A MIxed Strategy-Aware Model Integrating COMET for Emotional Support Conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2020. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5811–5820.

Qing Xu and Xiaohua Xuan. 2019. Nonlinear regression without i.i.d. assumption. *Probability, Uncertainty and Quantitative Risk*.

Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2019. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *arXiv:1911.02707*.

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 Asso- ciation for Computational Linguistics: ACL-IJCNLP*.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. Care: Commonsense-aware emotional response generation with latent concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14577–14585.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Pei Zhou, Pegah Jandaghi, Hyundong Cho, Bill Yuchen, Lin Jay Pujara, and Xiang Ren. 2021. Probing Commonsense Explanation in Dialogue Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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 (Volume 1: Long Papers)*, pages 1128–1137.

A The Details of Cognition Dynamics

Our goal is to calculate the effect of knowledge representation on the predictions of the linear transformation function $g_{\boldsymbol{\theta}}$ in the *workspace* module. The influence of excluding irrelevant knowledge representation can be interpreted as the change of $\boldsymbol{\theta}$ with respect to $\mathcal{L}_{\text{emo}}$, which is $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{emo}}(g_\theta(\mathbf{z}), g_\theta(\mathbf{z}_{\text{ctx}})), \mathbf{z} \in \mathbf{Z}$. Here, $\mathbf{Z} = \{\mathbf{Z}_r[0]\}, \mathbf{Z}_r[0] \in \mathbb{R}^d$. In order to eliminate the most irrelevant knowledge representation, we take the $\max(\cdot)$ on loss function with respect to $\mathbf{z} \in \mathbf{Z}$. However, it is challenging to calculate the gradient when we implement $\max(\cdot)$ on the groups of loss functions, because the above function is non-differentiable. Thus, we first bring differentiability for $\max_{\mathbf{z}\in\mathbf{Z}}\{\mathcal{L}_{\text{emo}}(g_\theta(\mathbf{z}), g_\theta(\mathbf{z}_{\text{ctx}}))\}$. To simplify notation, objective function is set as $\Phi(\boldsymbol{\theta}) = \max_{1\leq j\leq J} f_j(\boldsymbol{\theta})$, $f_j(\boldsymbol{\theta}) = \mathcal{L}_{\text{emo}}(g_\theta(\mathbf{z}_j), g_\theta(\mathbf{z}_{\text{ctx}}))$. Here, f_j denotes the loss function with respect to knowledge representation $\mathbf{z}_j$ and each f_j is differentiable. $g_{\boldsymbol{\theta}}$ is the parametric linear layer. Then, calculating the gradient of $\boldsymbol{\theta}$ turns into the following discrete mini-max problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \max_{1\leq j\leq J} f_j(\boldsymbol{\theta}). \tag{11}$$

In order to smooth objective function Φ during the iteration m, we linearize f_j at $\boldsymbol{\theta}_m$ and obtain the convex approximation of Φ as

$$\hat{\Phi}(\boldsymbol{\theta}) = \max_{1\leq j\leq J} \underbrace{\{f_j(\boldsymbol{\theta}_m) + \langle \nabla f_j(\boldsymbol{\theta}_m), \boldsymbol{\theta} - \boldsymbol{\theta}_m\rangle\}}_{\text{linearization term}}. \tag{12}$$

The linearization term smooths $\max(\cdot)$ function. Next step is to find descent direction, which minimizes $\hat{\Phi}$. However, $\hat{\Phi}$ is not strictly convex with respect to $\boldsymbol{\theta}$, the algorithm may not reach global minimum. So a regularization term $\|\boldsymbol{\theta} - \boldsymbol{\theta}_m\|_2$ is added for finding stable descent direction. Denote the descent direction $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}_m$, the discrete mini-max problem now is equivalent to

$$\min_{\boldsymbol{\delta},\nu} \quad \|\boldsymbol{\delta}\|_2 + \nu \tag{13a}$$

$$\text{s.t.} \quad f_j(\boldsymbol{\theta}_m) + \langle \nabla f_j(\boldsymbol{\theta}_m), \boldsymbol{\delta}\rangle \leq \nu,\ \forall 1 \leq j \leq J. \tag{13b}$$

Problem (13) is a semi-definite quadratic programming (QP) problem since we choose ℓ_2 norm as the regularization term. When the number of datapoints in subgroup is large, widely-used QP algorithms, such as active-set method, are time-consuming. Thus we turn to the dual problem. Consider the Lagrange multiplier for problem (13),

$$\begin{aligned} L(\boldsymbol{\delta}, \nu; \boldsymbol{\lambda}) = &\frac{1}{2}\|\boldsymbol{\delta}\|^2 + \nu \\ &+ \sum_{j=1}^{J} \lambda_j (f_j(\boldsymbol{\theta}_m) + \langle \nabla f_j(\boldsymbol{\theta}_m), \boldsymbol{\delta}\rangle - \nu). \end{aligned} \tag{14}$$

By strong duality theorem, the minimum of original problem is equal to the maximum of dual problem under specific constrains:

$$\min_{\boldsymbol{\delta},\nu} \max_{\boldsymbol{\lambda}\geq 0} L(\boldsymbol{\delta}, \nu; \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}\geq 0} \min_{\boldsymbol{\delta},\nu} L(\boldsymbol{\delta}, \nu; \boldsymbol{\lambda}) \tag{15}$$

Let $\mathbf{f} = (f_1, \cdots, f_J)^T$ and $\mathbf{G} = \nabla_{\boldsymbol{\theta}} \mathbf{f} \in \mathbb{R}^{d\times q}$. By setting $\mathbf{e} = \mathbf{1}$, the above problem is equivalent to

$$\max_{\boldsymbol{\lambda}\geq 0} \min_{\boldsymbol{\delta},\nu} \big(\frac{1}{2}\|\boldsymbol{\delta}\|^2 + \nu + \boldsymbol{\lambda}^T(\mathbf{f} + \mathbf{G}\boldsymbol{\delta} - \nu\mathbf{e})\big). \tag{16}$$

Note that

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{\delta}\|^2 + \nu + \boldsymbol{\lambda}^T(\mathbf{f} + \mathbf{G}\boldsymbol{\delta} - \nu\mathbf{e}) \\ &\quad = \frac{1}{2}\|\boldsymbol{\delta}\|^2 + \boldsymbol{\lambda}^T(\mathbf{f} + \mathbf{G}\boldsymbol{\delta}) + \nu(1 - \boldsymbol{\lambda}^T\mathbf{e}). \end{aligned} \tag{17}$$

If $1 - \boldsymbol{\lambda}^T\mathbf{e} \neq 0$, the objective function will be $-\infty$. Thus, we must have $1 - \boldsymbol{\lambda}^T\mathbf{e} = 0$ when the maximum is attained. The problem is converted to

$$\max_{\lambda_i\geq 0, \sum_{i=1}^{J}\lambda_i = 1} \min_{\boldsymbol{\delta}} \frac{1}{2}\|\boldsymbol{\delta}\|^2 + \boldsymbol{\lambda}^T\mathbf{G}\boldsymbol{\delta} + \boldsymbol{\lambda}^T\mathbf{f}. \tag{18}$$

Let the gradient of the inner minimization term to be zero, we have solution $\boldsymbol{\delta} = -\mathbf{G}^T\boldsymbol{\lambda}$. By changing the sign of (18), the maximization term is reduced to

$$\min_{\boldsymbol{\lambda}} \quad (\frac{1}{2}\boldsymbol{\lambda}^T\mathbf{G}\mathbf{G}^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\mathbf{f}) \tag{19a}$$

$$\text{s.t.} \quad \sum_{i=1}^{J} \lambda_i = 1, \lambda_i \geq 0. \tag{19b}$$

Suppose $\boldsymbol{\lambda}$ is the solution of the QP problem (13), then $\boldsymbol{\delta} = -\mathbf{G}^T\boldsymbol{\lambda}$ is the solution of problem above. Thus, we have the $\boldsymbol{\delta}$ as the change of eliminating irrelevant knowledge representation $\mathbf{z}$. By adding $\boldsymbol{\delta}$ to the refiner in decoder module, the final generated response would be less affected by the irrelevant knowledge. The effect of $\boldsymbol{\delta}$ is demonstrated by the generated responses in Table 5, and we also display how the elimination of irrelevant knowledge boost the performance.

Algorithm 1 Adaptive Knowledge Selection Method.

```
Input: The set of knowledge representation $\mathbf{Z} = \{\mathbf{Z}_r[0]\}, \mathbf{Z}_r[0] \in \mathbb{R}^d$, linear layer $g_\theta, \theta \in \mathbb{R}^{d\times q}$, the context representation vector $\mathbf{z}_{\text{ctx}} \in \mathbb{R}^d$ from dialogue history encoder, the objective function of emotion classification $\mathcal{L}_{\text{emo}}$.
  • Competition Stage:
  while $len(\mathbf{Z}) > 1$ do
      $m = 1$
      $I = \max_{\mathbf{z}\in\mathbf{Z}}\{\mathcal{L}_{\text{emo}}(g_\theta(\mathbf{z}), g_\theta(\mathbf{z}_{\text{ctx}}))\}$
      $\mathbf{f} = \{\mathcal{L}_{\text{emo}}(g_\theta(\mathbf{z}), g_\theta(\mathbf{z}_{\text{ctx}})), \mathbf{z} \in \mathbf{Z}\}$
      $\mathbf{G}_m = \nabla_{\boldsymbol{\theta}}\mathbf{f} \in \mathbb{R}^{d\times q}$
      Solve Lagrange multiplier $\boldsymbol{\lambda}$:
      $\min_{\boldsymbol{\lambda}} (\frac{1}{2}\boldsymbol{\lambda}^T\mathbf{G}_m\mathbf{G}_m^T\boldsymbol{\lambda} - \mathbf{f}^T\boldsymbol{\lambda})$
      $s.t. \quad \sum_{i=1}^{J}\lambda_i = 1, \lambda_i \geq 0.$
      if $m = 1$ then
          $\boldsymbol{\delta}_m = -\mathbf{G}_m^T\boldsymbol{\lambda}$
      else
          $\boldsymbol{\delta}_m = \boldsymbol{\delta}_{m-1} - \mathbf{G}_m^T\boldsymbol{\lambda}$
      end if
      $\mathbf{Z} = \mathbf{Z}_{-I}$
      $m = m + 1$
  end while
  • Broadcasting Stage:
  $\widetilde{\mathbf{h}}_k = \mathbf{h}_k + \boldsymbol{\delta}_m$
```

B Involved Existing Packages

Existing packages involved in this work include: 1) the open source codes, models weights and generated outcomes of Transformer (Vaswani et al., 2017), Multi-TRS (Rashkin et al., 2019), MoEL (Lin et al., 2019), MIME (Majumder et al., 2021), EmpDG (Li et al., 2021a), CEM (Sabour et al., 2021), and 2) the evaluation metrics from Natural Language Toolkit (Bird et al., 2009).

C Additional Case Study

We provide qualitative studies in Section 4.6. It includes 1) Ablation study of our cognition dynamics (Table 5); 2) Additional case study of generated responses from EmpatheticDiaglogues (Table 6); 3) Stackplot of the knowledge selection process for all the emotions in EmpatheticDiaglogues (Figure 5).

Context	Speaker: My friend borrowed quite a lot of money from me. I really do believe he'll repay me.
Knowledge	xIntent: To be helpful xWant: To repay the money xNeed: To ask for a loan xEffect: Gets a receipt xReact: Happy; Relieved
Emotion	***Trusting***
Selection Process	xReact → xEffect → xIntent → xNeed → xWant
Selected Knowledge	xWant: ***To repay the money***
Ours (w/o A*)	That is very nice of him. What did he do?
Ours (w/ A*)	I am sure he will repay you.
Context	Speaker: I had a nice meal and my favorite beverage after work.
Knowledge	xIntent: To relax after work xWant: To go to bed xNeed: To go to the restaurant xEffect: Has a full belly xReact: Satisfied; Happy
Content	***Trusting***
Selection Process	xNeed → xWant → xReact → xEffect → xIntent
Selected Knowledge	xIntent: ***To relax after work***
Ours (w/o A*)	What did you eat
Ours (w/ A*)	Sounds like a relaxing day. What did you drink?

Table 5: Ablation study of our cognition dynamics. $A*$ represents the adaptive knowledge selection method in the workspace module. Selection Process illustrates that at each the iteration, the most irrelevant knowledge is selected and eliminated.

Context	Speaker: My family and I are going on vacation in a few weeks. We rented a Largo Beachfront Condo and I can not wait! Listener: Oh I love the beach!! Which beach are you going to go to? Speaker: We are going to be staying at Miramar Beach in Florida, about an hour Panama City Beach. The ***view is amazing*** from our balcony and we are so ready!
MoEL **MIME** **EmpDG** **CEM**	That is great! I hope you have fun! That sounds like a lot of fun! That is awesome! I hope you are able to have a blast! That sounds like fun!
Selection Process **Selected Knowledge** **Emotion** **Ours**	xIntend → xEffect → xReact → xNeed → xWant xWant: ***To enjoy the view***; To enjoy the beach; To enjoy the day; To have fun ***Anticipating*** Sounds like a lot of fun. I have never been there, but I will check it out!
Gold for yourselfs is better.	I bet it is tough. A regular income is nice but the freedom you get from working
Context	Speaker: My parents keep taking my money that is not nice of them. Listener: Why do they keep taking it?? Speaker: They need it for other stuff. ***I never get a chance to save any of it***.
MoEL **MIME** **EmpDG** **CEM**	That is great! I hope you find a good job! That is a good thing to do. That is a very nice thing to do. Oh no! I am sorry to hear that.
Selection Process **Selected Knowledge** **Emotion** **Ours**	xReact → xEffect → xIntent → xWant → xNeed xNeed: ***To get rid of it***; To get rid of something; To buy it; To have it ***Annoyed*** I am sorry to hear that. Maybe you can talk to them about it.
Gold	Sorry to hear that. I found that it helps to have multiple savings accounts, so whenever you get a certain amount of money (like 5 , 10) save it and be strict not to spend it.
Context	Speaker: I went to visit a friend one day and her boyfriend made a pass at me. He actually suddenly kissed me, as I was waiting for my friend to return. I could not believe his nerve! Listener: Oh no, what did you do? Speaker: I quietly pushed him away, as I did not want my friend to hear. ***I was so shocked and upset***, since he knew I wouldn't want to make a scene and took advantage.
MoEL **MIME** **EmpDG** **CEM**	That is so nice of her. That is nice of you. That is a lot of people are not. I am sure you will be able to get another one. That is terrible. Did you find out who did it?
Selection Process **Selected Knowledge** **Emotion** **Ours**	xIntent → xEffect → xNeed → xWant → xReact xReact: ***Angry***; Upset; Sad; Annoyed; Scared ***Angry*** I would have done the same thing.
Gold	Did you ever tell your friend?
Context	Speaker: My mother recently went to the bank for me. I love her so much. Listener: That is really nice of her. It is nice to have someone to help you do things sometimes. Speaker: ***Nice people are the best***. Empathy is something we all need more of!
MoEL **MIME** **EmpDG** **CEM**	That is very nice of you. It is always nice to have family. That is a good thing to do. That is a very nice thing to have a good mom. That is nice of you.
Selection Process **Selected Knowledge** **Emotion** **Ours**	xNeed → xWant → xIntent → xReact → xEffect xEffect: ***Is loved***; Is respected; Is liked ***Caring*** I agree with you there. It is nice to have someone to lean on.
Gold	Yes. Too many people these days are mean to one another!

Table 6: Additional case study of generated responses from EmpatheticDiaglogues.

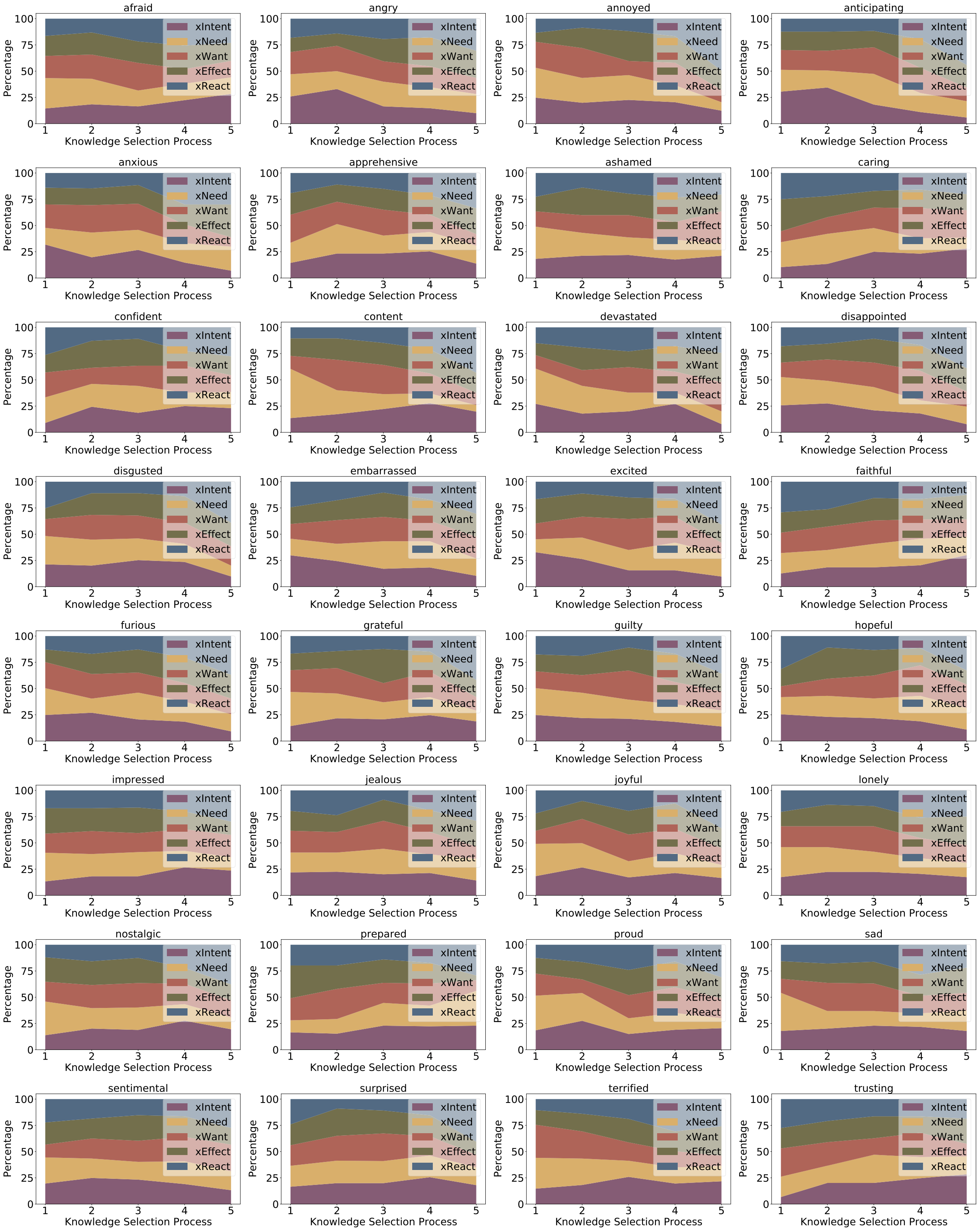

Figure 5: Stackplot of the knowledge selection process.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In section Limitations.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

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

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

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

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

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
4

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.