Towards Empathetic Dialogue Generation over Multi-type Knowledge

Qintong Li[1*], **Piji Li**[2], **Zhumin Chen**[1], **Zhaochun Ren**[1],
[1]School of Computer Science and Technology, Shandong University, Qingdao, China
[2]Tencent AI Lab, Shenzhen, China
qintongli@mail.sdu.edu.cn, pijili@tencent.com,
{chenzhumin, zhaochun.ren}@sdu.edu.cn

Abstract

Enabling the machines with empathetic abilities to provide context-consistent responses is crucial on both semantic and emotional levels. The task of empathetic dialogue generation is proposed to address this problem. However, lacking external knowledge makes it difficult to perceive implicit emotions from limited dialogue history. To address the above challenges, we propose to leverage multi-type knowledge, i.e, the commonsense knowledge and emotional lexicon, to explicitly understand and express emotions in empathetic dialogue generation. We first enrich the dialogue history by jointly interacting with two-type knowledge and construct an emotional context graph. Then we introduce a multi-type knowledge-aware context encoder to learn emotional context representations and distill emotional signals, which are the prerequisites to predicate emotions expressed in responses. Finally, we propose an emotional cross-attention mechanism to exploit the emotional dependencies between the emotional context graph and the target empathetic response. Conducted on a benchmark dataset, extensive experimental results show that our proposed framework outperforms state-of-the-art baselines in terms of automatic metrics and human evaluations.

Introduction

Studies on social psychology suggest that *empathy* is a crucial factor towards a more humanized dialogue system (Zech and Rimé 2005). Although plenty of researchers have attempted to control the emotional content of response either through an explicitly assigned emotional label (Zhou and Wang 2018; Zhou et al. 2018a; Wang and Wan 2018; Song et al. 2019; Shen and Feng 2020) or through a general term to encourage higher levels of affect (Asghar et al. 2018), it is still challenging for chatbots to conduct empathetic dialogue without the explicit emotion labels (empathetic dialogue problem) (Zhou et al. 2018a; Rashkin et al. 2019).

Several recent works have been proposed to address the empathetic dialogue problem based on multi-task learning (Rashkin et al. 2018, 2019; Wei et al. 2019; Lin et al. 2020) or mixture of experts (Lin et al. 2019). However, an unheeded deep concern is that humans usually rely on experience and external knowledge to acknowledge and express implicit emotions, which inspires us that external knowledge may play a crucial role in emotion understanding for empathetic dialogue systems. However, emotion perception and representation from external knowledge is still problematic for empathetic dialogue generation (Ghosal et al. 2019).

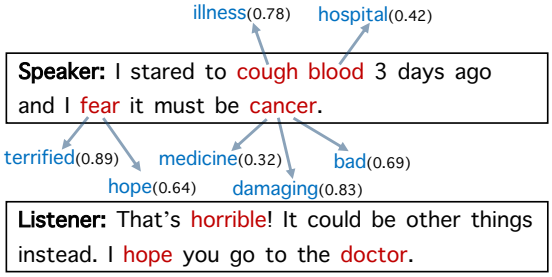

Figure 1: An example of empathetic dialogues with external knowledge from EMPATHETICDIALOGUES. Emotional-related words in the dialogue are highlighted in red color, whereas emotion-related concepts are marked in blue. Numbers in parentheses denote emotional intensity values.

According to a real-world example of empathetic dialogues (See Figure 1), we note that if we use non-stopwords of speaker's input as queries to acquire knowledge via two-type knowledge sources, i.e., a commonsense knowledge graph ConceptNet (Speer, Chin, and Havasi 2017) and an emotional lexicon NRC_VAD (Mohammad 2018), we can obtain several emotion-related concepts along with their emotional intensity values[1], such as "terrified (0.89)" and "bad (0.69)", which could easily help listener to respond in an empathetic way.

To illustrate this phenomenon concretely, we statistically investigate the functions of multi-type knowledge in emotional understanding on the dataset EMPATHETICDIALOGUES and the conclusions are illustrated in Figure 2. Figure 2(a) depicts that the response has almost NO non-stopword overlapping (0.5% of dialogue samples) with dialogue history. This phenomenon demonstrates that humans need to infer more knowledge to conduct empathetic dialogues. However, for a chatbot, perceiving emotions purely based on the dialogue history is tremendously difficult. Then what will happen if we incorporate multi-type knowledge into the system, say retrieve some emotion-related concepts

[*] This work was done when Qintong Li was an intern at Tencent AI Lab.

[1]Concepts and values come from different knowledge sources.

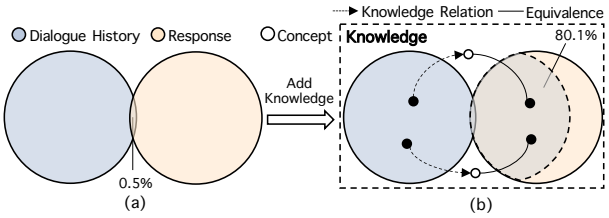

Figure 2: Relationships among dialogue history, responses, and knowledge.

by jointly considering both ConceptNet and NRC_VAD? The answer is depicted in Figure 2(b). We observe that, for most dialogue samples (80.1%) chatbots can directly obtain hints from the knowledge paths started by the non-stop tokens of the dialogue history. More importantly, with the knowledge from NRC_VAD, the emotion-related concepts have higher priorities in the concepts retrieved from ConceptNet, and thus are easier to be approached by empathetic dialogue models. Therefore, multi-type external knowledge is essential in acquiring useful emotional knowledge and improving the performance of empathetic dialogue generation.

Moreover, during the same investigations, we observe another phenomenon that emotional dependency and emotional inertia commonly appear with external knowledge in many real-world conversations. We label utterances with a CNN-based emotion classifier (Kim 2014), and visualize the emotion transitions from speakers to the listeners in Figure 3. In Figure 3, the darker diagonal grids show that listeners tend to mirror the emotion of their interlocutors to build rapport (Navarretta 2016). Moreover, there are also some complex emotional transition patterns besides the diagonal direction (in red frame). Therefore, intuitively, modelling emotional dependencies between interlocutors is crucial to enhance the accuracy of external knowledge representation in empathetic dialogues.

To this end, we propose a multi-type knowledge aware empathetic dialogue generation framework (**MK-EDG**). It consists of (1) an emotion-enhanced context graph constructed through the joint interaction between dialogue history and multi-type knowledge, (2) a multi-type knowledge-aware context encoder to emotion-enhanced context representations and distill fine-grained emotional signals, and (3) a multi-type knowledge-aware response generator conditioned on the context graph and distilled emotional signals. Conducted on the benchmark dataset EMPATHETICDIALOGUES, extensive experimental results demonstrate the effectiveness of MK-EDG in terms of both automatic metrics and human evaluation.

In summary, our contributions can be as follows:

• We leverage multi-type knowledge to enrich the dialogue history for empathetic dialogue generation, which is the early attempt. The proposed framework makes it easier to accurately perceive and appropriately express implicit emotions.

• We design a multi-type knowledge-aware context encoder and response generator to learn emotion-enhanced context representations and distill emotional signals, which are the prerequisites to learn the emotional dependencies between dialogue context and the target empathetic response.

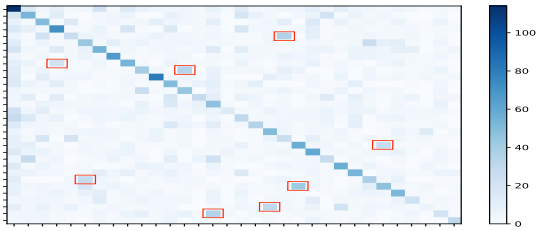

Figure 3: Emotion transition patterns. y-axis indicates the speaker's emotion label. x-axis indicates the listener's emotion label predicted by the classifier.

• Experimental results show the proposed framework outperforms competitive baselines in terms of both automatic evaluation metrics and human evaluations[2].

Related Work

Emotional Dialogue Generation: With the rise of data-driven learning approaches (Sutskever, Vinyals, and Le 2014; Vaswani et al. 2017), open-domain dialogue generation models have seen growing interests in recent years (Vinyals and Le 2015; Shang, Lu, and Li 2015; Serban et al. 2016; Li et al. 2016; Zhou et al. 2018b; Dinan et al. 2019). To control the emotional content of the target output, recent approaches generate emotional responses conditioning on a manually specified label (Zhou et al. 2018a; Li and Sun 2018; Zhou and Wang 2018; Huang et al. 2018; Colombo et al. 2019; Shen and Feng 2020). However, existing emotional dialogue models purely focus on whether the generated response matches a predetermined emotion, whereas in real-world scenarios the listener is capable to infer the emotion of the speaker (Rashkin et al. 2019).

Empathetic Dialogue Generation: Unlike the task of emotional dialogue generation, the task of empathetic dialogue generation avoids an additional step of determining which emotion type to respond explicitly (Skowron et al. 2013). Several works (Lubis et al. 2018; Rashkin et al. 2018; Zhong, Wang, and Miao 2019a; Wei et al. 2019; Shin et al. 2019; Chatterjee et al. 2019; Rashkin et al. 2019; Santhanam and Shaikh 2019; Lin et al. 2019, 2020; Zhong et al. 2020) have attempted to make dialogue models more empathetic. Rashkin et al. (2019) combined existing models in different ways to produce empathetic responses. Lin et al. (2019) softly combined the possible emotional responses from several separate experts to generate the final empathetic response. Besides the advancements in empathetic dialogue models, the emergence of new emotion-labelled dialogue corpora have also contributed to this research field. Li et al. (2017) introduced DAILYDIALOG dataset, with manually emotion labelling to each utterance in the multi-turn dialogue setting. Hsu et al. (2018) collected a multi-party dataset EMOTIONLINES where each utterance is labelled with one of seven emotions. However, only 5% of the utterances in DAILYDIALOG and 16.68% of the utterances in EMOTIONLINES have diverse emotional labels and others are either "none" or "happy" labels. Be-

[2]For the sake of fairness, we do not compare with models that use pre-trained language models, e.g., BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and GPT-2 (Radford et al. 2019).

cause of the extremely imbalanced data distribution, they are not suitable to be engaged as the benchmarks of empathetic dialogue generation. Rashkin et al. (2019) considered 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. In this work, we investigate how to leverage multi-type knowledge to explicitly improve the emotional understanding and expression in the task of empathetic dialogue generation on the dataset of EMPATHETICDIALOGUES.

Preliminaries

In this section, we introduce the two-type knowledge sources used in MK-EDG: the commonsense knowledge ConceptNet (Speer, Chin, and Havasi 2017) and the emotional lexicon NRC_VAD (Mohammad 2018).

ConceptNet is a large-scale knowledge graph that describes general human knowledge in natural language. It comprises 5.9M tuples, 3.1M concepts, and 38 relations. We denote each tuple (head concept, relation, tail concept, confidence score) as $\tau = (h, r, t, s)$ where the confidence score $s \in [1, 10]$. We use *min-max* normalization to scale s between 0 and 1:

$$\textit{min-max}(s) = \frac{s - \min_s}{\max_s - \min_s}, \tag{1}$$

where $\min_s = 1$ and $\max_s = 10$. For example, the confidence score s in tuple $\langle \text{birthday}, \text{RelatedTo}, \text{happy}, s \rangle$ is normalized from 2.69 to 0.19.

NRC_VAD is a lexicon of VAD (Valence-Arousal-Dominance) vectors with dimensions (V_a, A_r, D_o) for 20k English words. VAD vectors are culture-independent and widely adopted in Psychology (Mehrabian 1996). The interpretations of VAD vectors are presented in Table 1. For example, the VAD vector of word "nice" is: $[0.93, 0.442, 0.65]$.

Table 1: Interpretations of NRC_VAD vectors.

Dimensions	Values	Interpretations
Valence	$[0, 1]$	Negative - Positive
Arousal	$[0, 1]$	Calm - Excited
Dominance	$[0, 1]$	Submissive - Dominant

Inspired by Zhong, Wang, and Miao (2019b), we adopt NRC_VAD to compute emotion intensity values for words and concepts.

$$\eta(x_i) = \textit{min-max}(\left\| V_a(x_i) - \frac{1}{2}, \frac{A_r(x_i)}{2} \right\|_2), \tag{2}$$

where $\|.\|_k$ denotes L_k norm; $V_a(x_i)$ and $A_r(x_i)$ denote the values of valence and arousal dimensions in VAD vector of word x_i, respectively. If x_i not in NRC_VAD, $V_a(x_i)$ and $A_r(x_i)$ will be set to 0.5.

The Proposed MK-EDG Model

Overview

Formally, we are given a dialogue history of turns $[X_1, \ldots, X_M]$ where X_i is the i-th utterance, the commonsense knowledge ConceptNet, and the emotional knowledge NRC_VAD. The system needs to generate an empathetic response $Y = \{y_1, \ldots, y_n\}$ that is both emotionally appropriate and meaningful with the content.

The proposed framework MK-EDG is shown in Figure 4. Mk-EDG consists of three main parts:

• *Emotional Context Graph*: The dialogue history is enriched with bunches of emotion-related concepts by jointly interacting with two-type knowledge. The enriched dialogue history is represented as an emotion-enriched context graph G.

• *Multi-type Knowledge-aware Context Encoder*: It encodes the context graph into vector representations by a graph-aware Transformer encoder (Koncel-Kedziorski et al. 2019) and distills the emotional signals from the context graph, which is the prerequisite to generate the empathetic response.

• *Multi-type Knowledge-aware Response Generator*: It incorporates an emotional cross-attention mechanism to learn the emotional dependencies from the dialogue context graph and generate an empathetic response.

Emotional Context Graph

Following Lin et al. (2019), we flat dialogue history into a long token sequence S. For each non-stopword token $x_i \in S$, we first retrieve a set of candidate tuples $\left\{\tau_k^i = (x_i, r_k^i, c_k^i, s_k^i)\right\}_{k=1,\ldots,K}$ from ConceptNet. To refine the emotion-related knowledge from the commonsense knowledge, we introduce NRC_VAD and filter the retrieved tuples in two steps:

- **Relation Filtering**: r_k^i in ConceptNet are diversified and some are irrelevant, *e.g.,* "NotHasProperty". Therefore, we define an excluded relation list L_{ex} to filter the set of retrieved tuples and reserve the tuples whose $r_k^i \notin L_{ex}$ and $s_k^i > \alpha$, where α is a pre-defined threshold the same as the one in Zhong, Wang, and Miao (2019b).
- **Concept Ranking**: We derive a score f for each tuple τ_k^i according to three aspects (emotion, semantic, and confidence) simultaneously using knowledge from ConceptNet and NRC_VAD:

$$f\left(\tau_k^i\right) = \eta(c_k^i) + \cos(x_i, c_k^i) + s_k^i, \tag{3}$$

where $\eta(c_k^i)$ is the emotion intensity value of tail concept c_k^i, *min-max* is the min-max normalization function, cos is the cosine similarity function between tokens. We sort the candidate tuples in descending order of the scores and select top K' tuples $\left\{\tau_k^i = (x^i, r_k^i, c_k^i, s_k^i)\right\}_{k=1,\ldots,K'}$ as the emotional knowledge for each token x_i.

Finally, the dialogue history is enriched by emotional knowledge, i.e., K' tuples, which are inferred from two-type knowledge, and represented as the knowledge-enhanced context graph G. The tokens $\{x_i\}_{i=1,\ldots,|S|}$ in dialogue history, the tail concepts $\{c_k^i\}_{i=1,\ldots,|S|;k=1,\ldots,K'}$ in K' tuples, and an additional CLS token constitute the nodes $V = \{v_i\}_{i=1,\ldots,m}$ of G, where m is the number of nodes. G contains 3 directed relation types: (a) *sequence*: forward relation between two successive tokens, i.e., $x_i \rightarrow x_{i+1}$. (b) *emotion*: relation between the token x_i and its concepts c_k^i, i.e., $c_k^i \rightarrow x_i$. (c)

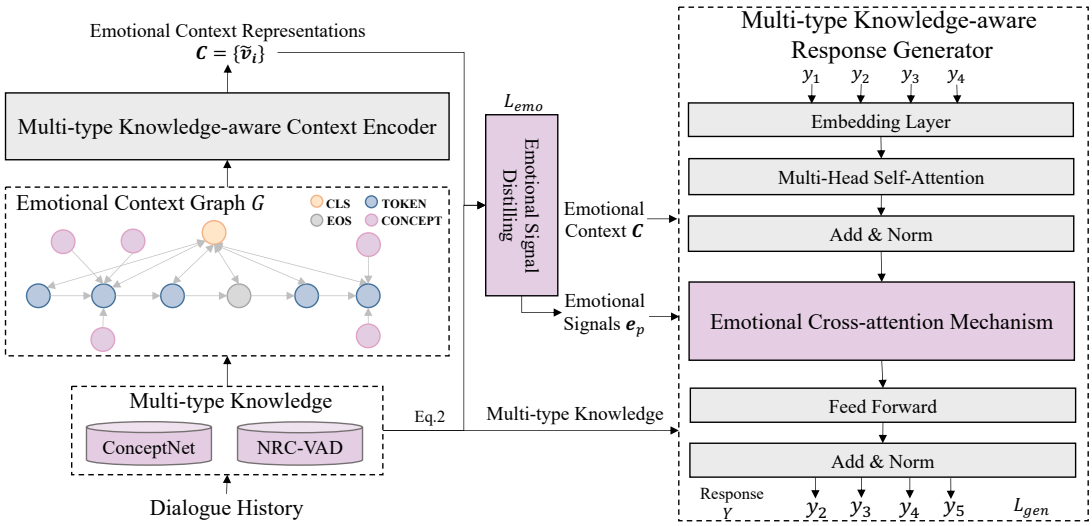

Figure 4: The overall architecture of MK-EDG.

globality: relation between CLS node and other nodes, i.e. $x_i \rightarrow$ CLS, $c_k^i \rightarrow$ CLS, and CLS $\rightarrow x_i$. All the three type relations among nodes are set to 1 in the adjacency matrix A of G.

Multi-type Knowledge aware Context Encoder

We describe the encoding details of the emotion-enriched context graph G and the procedure to distill emotional signals.

Context Graph Encoding We first use a word embedding layer and a positional embedding layer (Vaswani et al. 2017) to convert each node v_i into vectors $\mathbf{E}_w(v_i) \in \mathbb{R}^d$ and $\mathbf{E}_p(v_i) \in \mathbb{R}^d$, where d is the dimensionality of embeddings. In the multi-turn dialogue settings, distinguishing nodes in different utterances is helpful. So we incorporate the dialogue state embedding $\mathbf{E}_d(v_i)$ for node v_i. The vector representation of node v_i is the composition of three types of embeddings:

$$\mathbf{v}_i = \mathbf{E}_w(v_i) + \mathbf{E}_d(v_i) + \mathbf{E}_p(v_i). \tag{4}$$

Then we apply a multi-head graph-attention mechanism to update the node representations with emotional knowledge. Specifically, each node $\mathbf{v}_i$ is contextualized by attending to all its connected neighbour vertices $\{\mathbf{v}_j\}$.

$$\begin{aligned} \hat{\mathbf{v}}_i &= \mathbf{v}_i + \Big\|_{n=1}^{H} \sum_{j \in \mathcal{A}_i} \alpha_{ij}^n \mathbf{W}_v^n \mathbf{v}_j, \\ \alpha_{ij}^n &= a^n(\mathbf{v}_i, \mathbf{v}_j). \end{aligned} \tag{5}$$

Here, $\|$ denotes the concatenation of H attention heads, $\mathcal{A}_i$ denotes the neighborhood of v_i in the adjacency matrix A, a^n represent the self-attention mechanism of the n-th head in the following format:

$$a^n(\mathbf{q}_i, \mathbf{k}_j) = \frac{\exp((\mathbf{W}_q^n \mathbf{q}_i)^\top \mathbf{W}_k^n \mathbf{k}_j)}{\sum_{z \in \mathcal{A}_i} \exp((\mathbf{W}_k^n \mathbf{k}_z)^\top \mathbf{W}_q^n \mathbf{q}_i)}, \tag{6}$$

where $\mathbf{W}_q^n \in \mathbb{R}^{d_h \times d_h}$, $\mathbf{W}_k^n \in \mathbb{R}^{d_h \times d_h}$ are the linear transformations. $d_h = d/H$ is the dimension of each head.

Note that the previous operations are only conducted to the local context (immediate neighbours), thus we update the node representations with the global context information (all other nodes) to model the node interactions among different utterances. Precisely, we use transformer layers (Vaswani et al. 2017) to inject global information for all nodes $\{\hat{\mathbf{v}}_i\}_{i=1,\ldots,m}$.

$$\mathbf{h}_i^l = \text{LayerNorm}(\hat{\mathbf{v}}_i^{l-1} + \text{MHAtt}(\hat{\mathbf{v}}_i^{l-1})), \tag{7}$$

$$\hat{\mathbf{v}}_i^l = \text{LayerNorm}(\mathbf{h}_i^l + \text{FFN}(\mathbf{h}_i^l)), \tag{8}$$

where LayerNorm is the Layer Normalization trick proposed in (Ba, Kiros, and Hinton 2016); MHAtt is the multi-head self-attention sub-layer consiting of H attention heads; FFN is a two-layer feed-forward network with ReLU as hidden activation function. The transformer layers are stacked l times. The dialogue history is enriched by the emotional knowledge and represented as $\mathbf{C} = \{\tilde{\mathbf{v}}_i\}_{i=1,\ldots,m}$, where $\tilde{\mathbf{v}}_i = \hat{\mathbf{v}}_i^l$.

Emotional Signal Distilling Since there are no external emotional labels to be directly provided for MK-EDG, it needs to learn emotional signals from context graph to guide the empathetic response generation. The intuition is that humans can naturally pay extra attention to the emotional salient information during the conversation.

We utilize Eq. 2 to acquire knowledge, i.e., emotional intensity values $\{\eta_i\}$, from NRC_VAD for all nodes $\{\tilde{\mathbf{v}}_i\}_{i=1,\ldots,m}$. The emotion-enhanced context representation $\mathbf{c}_e$ can be derived by the weighted summation of all the node representations:

$$\mathbf{c}_e = \sum_{i=1}^{m} \frac{\exp(\eta_i)}{\sum_j \exp(\eta_j)} \tilde{\mathbf{v}}_i. \tag{9}$$

Then a linear layer with $\mathrm{softmax}$ operation projects the vector $\mathbf{c}_e$ into an emotion category distribution P_e over the emotion label e to identify the emotion signal for the target response:

$$\mathbf{e}_p = \mathbf{W}_e \mathbf{c}_e + b_e, \tag{10}$$

$$P_e(e|\mathbf{C}) = \text{softmax}(\mathbf{e}_p), \tag{11}$$

where $\mathbf{W}_e \in \mathbb{R}^{q\times d}$, $b_e \in \mathbb{R}^q$, and q is the number of emotion categories. During training, we employ negative log-likelihood as the loss function to conduct the parameter learning:

$$\mathcal{L}_{emo} = -\log(P_e(e^*|\mathbf{C})), \tag{12}$$

where e^* denotes the ground truth emotional label of dialogue history. The obtained intermediate emotional vector $\mathbf{e}_p$ will be fed into the decoder as a crucial emotional signal to guild the empathetic response generation.

Multi-type Knowledge aware Response Generator

The emotion vector $\mathbf{e}_p \in \mathbb{R}^{1\times q}$ is firstly be transformed by a linear transformation into $\mathbf{e}'_p \in \mathbb{R}^{1\times d}$. Then $\mathbf{e}'_p$ is concatenated with the embeddings of the decoder input tokens $y_{1:j-1}$ into representations $\mathbf{Y}_{emb} = \{\mathbf{y}_t\}_{t=0,\ldots,j-1}$ where $\mathbf{y}_0 = \mathbf{e}'_p$. We feed $\mathbf{Y}_{emb}$ into the response generator.

The generator is built based on L Transformer layers as well. For each Transformer decoder layer, the decoder inputs $\mathbf{Y}_{emb}$ are updated into new vector representations $\mathbf{Y} = \{\mathbf{y}_t\}_{t=0,\ldots,j-1}$.

To incorporate the emotional signals $\mathbf{c}_e$ into the generation process and help the generator to produce more appropriate emotions, an **Emotional Cross-attention Mechanism** E-CatM is designed to replace the original cross-attention sub-layer in the Transformer decoder layer. With the emotion dependency vectors produced by E-CatM, the representations of the predicted response are enhanced as follows:

$$\mathbf{D} = \mathbf{Y} + \mathbf{W}_m\text{E-CatM}(\mathbf{Y}, \mathbf{C}), \tag{13}$$

where E-CatM is the concatenation of the context vector from cross multi-head attention CMHAtt (Vaswani et al. 2017) and emotion vector $\mathbf{c}_e$:

$$\text{E-CatM}(\mathbf{Y}, \mathbf{C}) = [\text{CMHAtt}(\mathbf{Y}, \mathbf{C})\|\mathbf{c}_e]. \tag{14}$$

Similar with the Transformer decoder, the remaining operations are as following:

$$\hat{\mathbf{D}} = \text{LayerNorm}(\mathbf{D}), \tag{15}$$

$$\hat{\mathbf{Y}} = \text{LayerNorm}(\hat{\mathbf{D}} + \text{FFN}(\hat{\mathbf{D}})), \tag{16}$$

where $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_j]$. Then the response generator yields the distribution over the vocabulary for the next j-th token:

$$\alpha^g(y_j|\mathbf{C}, y_{<j}) = \text{softmax}(\mathbf{W}_o\hat{\mathbf{y}}_\mathbf{j}). \tag{17}$$

Since external concepts can introduce the emotional knowledge for empathetic responding, we compute a probability p_g of copying from nodes $\{v_i\}_{i=1,\ldots,m}$ in the context graph in a manner similar to See, Liu, and Manning (2017) and derive the final next-token probability distribution (y_j):

$$p_g = \sigma(\mathbf{W}_c\hat{\mathbf{y}}_j + b_c), \tag{18}$$

$$p(y_j) = (1 - p_g) * \alpha^c + p_g * \alpha^g, \tag{19}$$

where the copy probability α^c over the knowledge-enriched dialogue context is obtained from the concatenation of attention weights $\{\alpha^c_i = \alpha(\hat{\mathbf{y}}_j, \tilde{\mathbf{v}}_i)$, for $\tilde{\mathbf{v}}_i \in \mathbf{C}\}$. The computation formation of function α is same in (See, Liu, and Manning 2017).

As most dialogue generation tasks, we use standard maximum likelihood estimator (MLE) as the optimization objective:

$$\mathcal{L}_{gen} = -\log p(y_j|y_{<j}, \mathbf{C}). \tag{20}$$

Finally, considering all the aforementioned components, we define a joint loss function in the following equation:

$$\mathcal{L} = \gamma_1\mathcal{L}_{emo} + \gamma_2\mathcal{L}_{gen}, \tag{21}$$

where γ_1, γ_2 are hyper-parameters that control the weights of the two losses. We set $\gamma_1 = \gamma_2 = 1$. All the parameters are jointly trained in an end-to-end paradigm.

Experimental Settings

Data Preparation

We conduct our experiments on the EMPATHETICDIALOGUES dataset (Rashkin et al. 2019). EMPATHETICDIALOGUES is a large-scale multi-turn empathetic dialogue dataset collected on the Amazon Mechanical Turk, containing about 25k one-to-one open-domain conversation. Specifically, Rashkin et al. (2019) pair two crowd-workers: a speaker and a listener. The speaker is asked to talk about the personal emotional feelings. The listener infers the underlying emotion through what the speaker says and responds empathetically. The dataset provides 32 evenly distributed emotion labels. At training time, the emotional label of the historical utterances (*i.e.*, the speaker) acts as a supervised signal, while we hide the label in test time to evaluate the empathetic ability of all the models. We treat the historical utterances as the system input and the listener's response as the target output. Then we obtain 17,802 dialogues in the training set, 2,628 in the validation set, and 2,494 in the testing set. The average historical utterances and response lengths are 2.1 utterances and 13.5 tokens respectively.

Baselines

We compare with the state-of-the-art baselines as follows: (1) **Transformer** (Vaswani et al. 2017): A Transformer-based sequence to sequence model that is trained based on MLE loss. (2) **EmoPrepend-1** (Rashkin et al. 2019): An extension of the Transformer model which incorporates an additional supervised emotion classifier. The whole model is jointly trained by optimizing both the classification and generation loss. (3) **MoEL** (Lin et al. 2019): Another extension of Transformer model which softly combines the response representations from different transformer decoders. Each decoder is optimized to focus on one type of emotion accordingly.

Additionally, we also conduct ablation studies to better analyze the influence of different components in our model: (1) **w/o MKCE**: The MK-EDG model without the Multi-type Knowledge-aware Context Encoder. (2) **w/o E-CatM**: The MK-EDG model without the Emotional Cross-attention Mechanism.

Since both the two-type knowledge sources are utilized in the multi-type knowledge-aware context encoder and response generator, we do not conduct experiments to analyze the effects of single-type knowledge.

Table 2: Evaluation results of baselines and our models. **Bold face** indicates leading results in terms of the corresponding metric. t-test is conducted between MK-EDG and baselines. (underline: *p-value* < 0.05, *: p-value < 0.01)

Models	Accuracy	Perplexity	Distinct-1	Distinct-2	Empathy	Relevance	Fluency
Transformer	-	37.73	0.47*	2.04*	3.11	3.47	3.66
EmoPrepend-1	0.3452	38.30	0.46*	2.08*	3.23	3.51	**3.74**
MoEL	0.3524	38.04	0.44*	2.10*	3.37	3.78	3.64
MK-EDG	**0.3931**	**34.85**	**1.48**	**4.90**	**3.49**	**3.91**	3.65

Implementation Details

We lowercase the characters, tokenize the sequences and retain a vocabulary with 24,646 tokens. We use pre-trained Glove vectors (Pennington, Socher, and Manning 2014) to initialize the word embedding. All common hyperparameters are the same as the work in (Lin et al. 2019). The maximum introducing numbers of external concepts per dialogue and per token are set as 10 and 5, respectively. The threshold α used in dialogue context graph construction is 0.1. We implemented all models in PyTorch (Paszke et al. 2017) with a single Tesla V100 GPU, and train models using Adam optimization (Kingma and Ba 2014) with a mini-batch size of 16. The parameters of the Transformer, Emoprepend_1, MoEL, and our model Know_EDG are 16M, 16M, 21M, and 31M. We varied the learning rate during training following Vaswani et al. (2017). Early stopping is applied when training. The training time of model Know_EDG is 3 hours for around 21000 iterations. When inference, we set the maximum decoding step as 30.

Evaluation Metrics

Automatic Evaluations To evaluate the model at the emotional level, we adopt *Emotion Accuracy* as the agreement between the ground truth emotion labels and the predicted emotion labels by the emotion identifier. Liu et al. (2016) have verified BLEU is not suitable for measuring dialogue generation quality due to its low correlation with human judgment; METEOR (Banerjee and Lavie 2005) and ROUGE (Lin 2004) have the same problem. Therefore, following previous emotion-related studies (Zhou et al. 2018a; Rashkin et al. 2019; Song et al. 2019; Wei et al. 2019), we adopt Perplexity (Serban et al. 2015), Distinct-1, and Distinct-2 (Li et al. 2014) to evaluate comparisons in our experiments: Perplexity measures the high-level general quality of the generation model. Distinct-1 / Distinct-2 is the proportion of the distinct unigrams / bigrams in all the generated results to indicate the diversity.

Human Evaluations. We randomly sample 100 dialogues and their corresponding generations from our model as well as the baselines. We recruit three professional annotators from a third-party company to evaluate the responses generated by different models. All models are evaluated in terms of following 3 metrics: Empathy, Relevance and Fluency (Rashkin et al. 2019; Lin et al. 2019). Empathy measures whether the listener's responses show the understanding of the speaker's feelings; Relevance evaluates whether the generated responses are on-topic with the historical utterances; Fluency measures the grammatical correctness and readability of the generated responses. Each metric is rated on five-scale, where 1, 3, and 5 indicate unacceptable, moderate, and excellent performance, respectively.

Table 3: Ablation study.

Models	Accuracy	Perplexity	Distinct-1	Distinct-2
MK-EDG	**0.3931**	**34.85**	1.48	4.90
w/o MKCE	0.3637	35.52	**1.55**	**5.61**
w/o E-CatM	0.3802	35.48	0.95	2.69

Table 4: Result of human A/B test. Tests are conducted pairwise between MK-EDG and baseline models.

Models	Win	Loss	Tie
MK-EDG vs Transformer	43.8%	17.5%	38.7%
MK-EDG vs EmoP	38.3%	18.0%	43.7%
MK-EDG vs MoEL	36.6%	20.6%	42.8%

Results and Analysis

Automatic Evaluation Results In Table 2, we observe that our model MK-EDG outperforms state-of-the-art baseline MoEL by a large margin in terms of all automatic metrics. The noticeable improvement indicates the effectiveness of our knowledge-enhanced framework in empathetic expression and response diversity. EmpPrepend-1 and MoEL have similar performance, as both of them only use the historical utterances to infer emotional states and generate responses. Without emotion modelling, Transformer only generates fluent responses based on semantic mapping, but fail to express diverse responses.

We also perform an ablation study for better understanding the contributions of the main parts of our model. As shown in Table 3, after we remove the multi-type knowledge-aware (w/o MKCE model), both the emotion accuracy and perplexity performance become obviously worse, indicating that injecting external knowledge is consistently critical for emotion understanding and model generation quality. We also investigate the effect of removing emotional cross-attention mechanism (i.e., w/o E-CatM model). We notice that the scores decrease dramatically on all metrics, which demonstrates the effectiveness of the Emotion-Focused Attention Mechanism in modelling emotional dependencies.

Human Evaluation Results Table 2 illustrates that MK-EDG obtains the best performance on both Empathy and

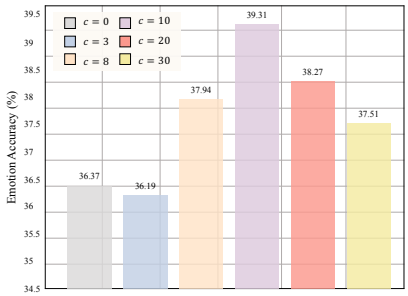

Figure 5: Emotion accuracy with respect to the number of external concepts. $c = 8$ indicates at most we introduce 8 external concepts for each dialogue.

Table 5: The visualization of the cross-attention weights in MoEL and MK-EDG.

Utterance	It inspires me to try and do something to keep healthy every day .
MoEL	I am sure they will be able to have a good time.
Utterance	It inspires me to try and do something to keep healthy every day .
Knowledge	effort , fight , good , life , raise , grow , protect , health
MK-EDG	I can not wait to try to get a little makes me feel better.

Relevance scores. This suggests that the emotion-focused attention mechanism on knowledge-enriched dialogue context helps to capture implicit emotions, improve the topic consistency, and elicits a more appropriate response. We see there is no obvious difference among models in terms of Fluency. We deduce it's because the generated responses by Transformer are already fluent and grammatical. Additionally, we carried out pairwise response comparison to directly compare the dialogue quality gains in Table 4. The results confirm that the responses from MK-EDG are more preferred by human judges.

External Multi-type Knowledge Analysis

To further investigate the impact of external multi-type knowledge, we train our model with different number of concepts and evaluate in terms of Accuracy as shown in Figure 5. With increasing the number of concepts, the performance is rising. However, if we introduce too many concepts, the accuracy no longer increase or even decrease. Therefore, external knowledge is more suitable to be the auxiliary information to understand the emotional states in the historical utterances.

Emotional Cross-attention Analysis

Table 5 shows an example illustrating the cross-attention weights of the dialogue context. Baseline MoEL put the major attention on the general words, which leads to a context-inconsistent and emotion-inappropriate response. In comparison, the MK-EDG model puts the highest attention probability on the words containing informative meaning, e.g., "fight" and "grow" in external knowledge and "keep" and "healthy" in historical utterances. We can conclude that the proposed emotion-focused attention mechanism can teach the model to generate responses from meaningful and emotional words.

Table 6: Generated responses from Transformer (short in Transfmr), EmoPrepend-1 (short in EmoP), MoEL, and MK-EDG (short in EDG) in two different speaker's emotion states. Tokens in underline represent knowledge-related words.

Emo.	**Terrified**
Cont.	X_1: Do you know how crazy it is to skydive? X_2: I have a fear of falling from high places. X_3: It gave me the biggest rush that is for sure.
Gold	**I think I would pass out from fear lol.**
Transfmr	I am sure it was.
EmoP	I am sure it was a lot of fun.
MoEL	I am sure you will do it!
Know.	danger, pain, travel, scared, excited, furious
EDG	I would have been so scared.
Emo.	**Surprised**
Cont.	X_1: I realized yesterday I was pregnant, I was in shock!
Gold	**Woah, that is huge news. How are you handling the news.**
Transfmr	Oh no! I am so sorry to hear that.
EmoP	Oh no! Did you get a job?
MoEL	That is so awesome! I am glad to hear that.
Know.	experience, surprise, mother, pain, feeling
EDG	Oh wow ! that is a great feeling.

Case Study

Cases from MK-EDG and baseline models are listed in Table 6. In the first case, MK-EDG generates both coherent and informative responses with a proper negative emotion by replying with "scared". However, without emotional knowledge, all baselines fail to recognize the negative emotion, i.e., terrified. In the second case, MK-EDG model generates the most context-consisten response, which contains context-related word ("feeling") and emotion-rated word ("Oh wow"). Both the two cases show that the MK-EDG model can balance the performances between content and emotion.

Conclusion

We propose a multi-type knowledge aware empathetic dialogue generation framework, MK-EDG, to enhance the performance of empathetic dialogue generation. We enrich the dialogue utterances by jointly interacting with multi-type knowledge and constructing an emotional context graph. Then we employ a multi-type knowledge-aware context encoder to learn emotional context representations and distill the emotional signals. Finally, we design an emotional cross-attention mechanism to exploit the emotional dependencies between the emotional context graph and the target empathetic response. Experimental results show that our framework outperforms all state-of-the-art baselines in terms of automatic metrics and human evaluations.

Ethics Statement

To comply with the ethics policy in AAAI 2021, we analyze the potential ethical impact of our work, including transparency, privacy, and politeness. In this paper, we focus on empathetic dialogue generation where an empathetic chatbot recognizes feelings in the conversation partner and replies accordingly. We identify several potential ethical concerns during empathetic dialogue generation as follows:

Transparency: The motivation of our work is reacting to emotional cues and displaying a caring attitude, where the empathetic chatbot predicates an emotional signal as the guidance for the target empathetic response. MK-EDG provides the faithful and trustworthy service in two aspects: (1) it facilitates the model's ability to understand of users feelings and emotional experiences; (2) it explicitly details the model's empathy behaviour in a way the conversation partner can understand.

Privacy: The dataset we used is collected by (Rashkin et al. 2019) using Amazon Mechanical Turk, therefore, it does not harm the privacy of real users. If we apply MK-EDG in practice, we will protect the privacy of users by following a two-step masking process (Hsu et al. 2018).

Politeness: Previous models (Ritter, Cherry, and Dolan 2010; Mazaré et al. 2018; Radford et al. 2019) trained on vast amounts of barely curated text scrapes or social media conversations could exhibit some of the aggressive and callous responses that have been observed in spontaneous internet conversations (Rashkin et al. 2019). The dataset EMPATHETICDIALOGUES used by MK-EDG is collected via crowdsourcing which can relieve the impolite behaviour. Moreover, the success of ELIZA also demonstrates that both intelligence and empathy are the important characteristics for the chatbot to engage users (Ruane, Birhane, and Ventresque 2019).

Given above detailed demonstrations, we believe our research work will not violate the AAAI Publications Ethics and Malpractice Statement.

References

Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; and Mou, L. 2018. Affective neural response generation. In *ECIR*, volume 10772 of *Lecture Notes in Computer Science*, 154–166. Springer.

Ba, L. J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *CoRR* abs/1607.06450.

Banerjee, S., and Lavie, A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, 65–72. Association for Computational Linguistics.

Chatterjee, A.; Gupta, U.; Chinnakotla, M. K.; Srikanth, R.; Galley, M.; and Agrawal, P. 2019. Understanding emotions in text using deep learning and big data. *Comput. Hum. Behav.* 93:309–317.

Colombo, P.; Witon, W.; Modi, A.; Kennedy, J.; and Kapadia, M. 2019. Affect-driven dialog generation. In *NAACL*, 3734–3743.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*. OpenReview.net.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Hsu, C.; Chen, S.; Kuo, C.; Huang, T. K.; and Ku, L. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *LREC*. European Language Resources Association (ELRA).

Huang, C.; Zaiane, O.; Trabelsi, A.; and Dziri, N. 2018. Automatic dialogue generation with expressed emotions. In *NAACL*, 49–54.

Kim, Y. 2014. Convolutional neural networks for sentence classification. In *EMNLP*, 1746–1751.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koncel-Kedziorski, R.; Bekal, D.; Luan, Y.; Lapata, M.; and Hajishirzi, H. 2019. Text generation from knowledge graphs with graph transformers. In *NAACL*, 2284–2293.

Li, J., and Sun, X. 2018. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. In *EMNLP*, 678–683.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2014. A diversity-promoting objective function for neural conversation models. In *NAACL*, 110–119.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016. Deep reinforcement learning for dialogue generation. In *EMNLP*, 1192–1202. The Association for Computational Linguistics.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*, 986–995.

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 2019. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*.

Lin, Z.; Xu, P.; Winata, G. I.; Siddique, F. B.; Liu, Z.; Shin, J.; and Fung, P. 2020. Caire: An end-to-end empathetic chatbot. In *AAAI*, 13622–13623.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, C.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2122–2132. The Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*.

Lubis, N.; Sakti, S.; Yoshino, K.; and Nakamura, S. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *AAAI*.

Mazaré, P.-E.; Humeau, S.; Raison, M.; and Bordes, A. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Mehrabian, A. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology* 14(4):261–292.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *ACL*, 174–184.

Navarretta, C. 2016. Mirroring facial expressions and emotions in dyadic conversations. In *LREC*.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y. 2018. I know the feeling: Learning to converse with empathy. *CoRR* abs/1811.00207.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, 5370–5381.

Ritter, A.; Cherry, C.; and Dolan, B. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 172–180.

Ruane, E.; Birhane, A.; and Ventresque, A. 2019. Conversational ai: Social and ethical considerations. In *AICS*, 104–115.

Santhanam, S., and Shaikh, S. 2019. Emotional neural language generation grounded in situational contexts. *CoRR* abs/1911.11161.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*, 1073–1083. Association for Computational Linguistics.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2015. Hierarchical neural network generative models for movie dialogues. *CoRR* abs/1507.04808.

Serban, I. V.; Lowe, R.; Charlin, L.; and Pineau, J. 2016. Generative deep neural networks for dialogue: A short review. *CoRR* abs/1611.06216.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural responding machine for short-text conversation. In *ACL*, 1577–1586.

Shen, L., and Feng, Y. 2020. CDL: curriculum dual learning for emotion-controllable response generation. *CoRR* abs/2005.00329.

Shin, J.; Xu, P.; Madotto, A.; and Fung, P. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *CoRR* abs/1906.08487.

Skowron, M.; Theunis, M.; Rank, S.; and Kappas, A. 2013. Affect and social processes in online communication–experiments with an affective dialog system. *IEEE Transactions on Affective Computing* 4(3):267–279.

Song, Z.; Zheng, X.; Liu, L.; Xu, M.; and Huang, X.-J. 2019. Generating responses with a specific emotion in dialog. In *ACL*, 3685–3695.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010. Curran Associates Inc.

Vinyals, O., and Le, Q. V. 2015. A neural conversational model. *CoRR* abs/1506.05869.

Wang, K., and Wan, X. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, 4446–4452. ijcai.org.

Wei, W.; Liu, J.; Mao, X.; Guo, G.; Zhu, F.; Zhou, P.; and Hu, Y. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *CIKM*, 1401–1410. ACM.

Zech, E., and Rimé, B. 2005. Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice* 12(4):270–287.

Zhong, P.; Sun, Y.; Liu, Y.; Zhang, C.; Wang, H.; Nie, Z.; and Miao, C. 2020. Endowing empathetic dialogue systems with personas. *CoRR* abs/2004.12316.

Zhong, P.; Wang, D.; and Miao, C. 2019a. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *AAAI*, volume 33, 7492–7500.

Zhong, P.; Wang, D.; and Miao, C. 2019b. Knowledge-enriched transformer for emotion detection in textual conversations. In *EMNLP-IJCNLP*, 165–176.

Zhou, X., and Wang, W. Y. 2018. Mojitalk: Generating emotional responses at scale. In *ACL*, 1128–1137.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, 4623–4629.