Knowledge Bridging for Empathetic Dialogue Generation

Qintong Li[1,3], Piji Li[2*], Zhaochun Ren[1*], Pengjie Ren[1], Zhumin Chen[1]

[1]School of Computer Science and Technology, Shandong University, Qingdao, China
[2]Tencent AI Lab, Shenzhen, China
[3]Department of Computer Science, The University of Hong Kong, Hong Kong SAR, China
qtleo@outlook.com, {zhaochun.ren, chenzhumin}@sdu.edu.cn, lipiji.pz@gmail.com, jay.ren@outlook.com

Abstract

Lack of external knowledge makes empathetic dialogue systems difficult to perceive implicit emotions and learn emotional interactions from limited dialogue history. To address the above problems, we propose to leverage external knowledge, including commonsense knowledge and emotional lexical knowledge, to explicitly understand and express emotions in empathetic dialogue generation. We first enrich the dialogue history by jointly interacting with external knowledge and construct an emotional context graph. Then we learn emotional context representations from the knowledge-enriched emotional context graph and distill emotional signals, which are the prerequisites to predicate emotions expressed in responses. Finally, to generate the empathetic response, we propose an emotional cross-attention mechanism to learn the emotional dependencies from the emotional context graph. Extensive experiments conducted on a benchmark dataset verify the effectiveness of the proposed method. In addition, we find the performance of our method can be further improved by integrating with a pre-trained model that works orthogonally.

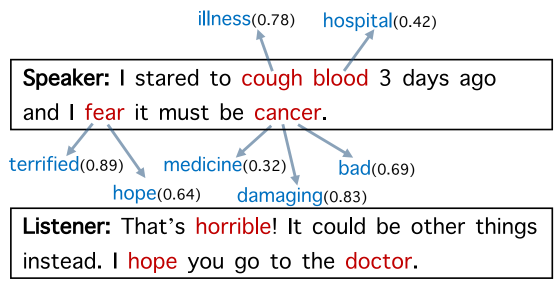

Figure 1: An example of empathetic dialogues with external knowledge from EMPATHETICDIALOGUES (Rashkin et al. 2019). Emotion-related words in the dialogue are highlighted in red color, whereas emotion-related concepts are marked in blue. Numbers in parentheses denote emotional intensity values.

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 dialogues 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), the mixture of experts (Lin et al. 2019), emotion mimicry (Majumder et al. 2020), or multi-resolution user feedback (Li et al. 2020).

However, an unheeded deep concern is that humans usually rely on experience and external knowledge to acknowledge and express implicit emotions (Zhong, Wang, and Miao 2019b). Figure 1 shows a real-world example of empathetic dialogues. If we use non-stopwords of speaker's input as queries to acquire knowledge via external knowledge, we can obtain various emotion-related concepts along with their emotional intensity values, which play a crucial role in emotion understanding for empathetic dialogue systems.

To exploit this phenomenon more concretely, we quantitatively investigate effects of external knowledge in understanding emotions on an empathetic dialogue corpus, i.e., EMPATHETICDIALOGUES (Rashkin et al. 2019). Figure 2(a) depicts that the response has almost NO non-stopword overlapping (0.5% of dialogue samples) with the dialogue history. This phenomenon implies that humans need to infer more external knowledge to conduct empathetic dialogues. By contrast, if we incorporate external knowledge (i.e., emotion-related concepts) into the system, 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 (shown in Figure 2(b)). Hence, external knowledge is essential in acquiring useful emotional knowledge and improving the performance of empathetic dialogue generation. However, emotion perception and representation from external knowledge is still problematic for empathetic dialogue generation.

During the investigations, we observe another phenomenon that emotional dependency and emotional inertia commonly

*Corresponding authors.

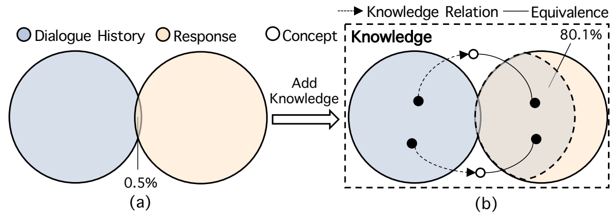

Figure 2: Relationships among the dialogue history, responses, and external knowledge.

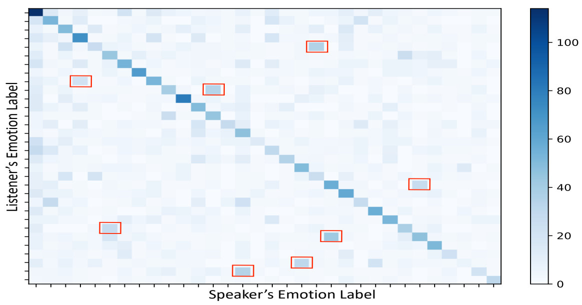

Figure 3: Emotion transition patterns.

appear with external knowledge in empathetic 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, it is crucial to model emotional dependencies between interlocutors.

To this end, we propose a **K**nowledge-aware **EMP**athetic dialogue generation method (**KEMP**). It consists of three components: an emotional context graph, an emotional context encoder, and an emotion-dependency decoder. The emotional context graph is constructed via integrating the dialogue history with external knowledge. The emotional context encoder employs the graph-aware transformer to learn the graph embeddings, and propose an emotional signal perception procedure to perceive context emotions that lead the response generation. Conditioned on the knowledge-enriched context graph, the emotion-dependency decoder particularly models emotion dependencies to generate empathetic response. A multi-task learning framework is applied to jointly optimize our objectives.

Conducted on the benchmark dataset EMPATHETICDIALOGUES (Rashkin et al. 2019), extensive experimental results demonstrate the effectiveness of KEMP in terms of both automatic and human evaluations.

In summary, our contributions are as follows: (a) We propose KEMP which is able to accurately perceive and appropriately express implicit emotions. To the best of our knowledge, this is the first attempt to leverage external knowledge to enhance empathetic dialogue generation. (b) We design an emotional context encoder and an emotion-dependency decoder to learn the emotional dependencies between the emotion-enhanced representations of the dialogue history and target response. (c) We conduct extensive experiments and analyses to demonstrate the effectiveness of KEMP.[1]

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. 2016b; 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; Wei et al. 2019; 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 (Rashkin et al. 2018; Zhong, Wang, and Miao 2019a; 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; Majumder et al. 2020; Li et al. 2020) have attempted to make dialogue models more empathetic. Rashkin et al. (2019) combine existing models in different ways to produce empathetic responses. Lin et al. (2019) softly combine the possible emotional responses from several separate experts. Majumder et al. (2020) considere of this polarity-based emotion clusters and emotional mimicry. Li et al. (2020) propose a multi-resolution adversarial framework which considers multi-granularity emotion factors and users' feedback.

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; Hsu et al. 2018; Rashkin et al. 2019). 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. In this work, we investigate how to leverage external knowledge to explicitly improve the emotional understanding and expression in the task of empathetic dialogue generation on the dataset of EMPATHETICDIALOGUES.

Preliminaries

In this work, external knowledge serves as the bridge to improve emotion perception and emotion expression capabilities. Therefore, we first introduce the two-type knowledge

[1] Code and dataset are available at http://github.com/qtli/KEMP.

sources used in KEMP: 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, playing an effective role in sentiment-related task (Ghosal et al. 2020). It comprises 5.9M tuples, 3.1M concepts, and 38 relations. We denote each tuple (head concept, relation, tail concept, confidence score) as $\tau = (x, r, c, s)$, e.g., $\langle \mathrm{birthday}, \mathrm{RelatedTo}, \mathrm{happy}, 0.19 \rangle$.

NRC_VAD is a lexicon of VAD (Valence-Arousal-Dominance) vectors with 3-dimensions (V_a, A_r, D_o) for 20k English words, e.g., the VAD vector of word "nice" is: $[0.93, 0.442, 0.65]$. VAD vectors are culture-independent and widely adopted in Psychology (Mehrabian 1996). The interpretations of VAD vectors are presented in Table 1.

Table 1: Interpretations of VAD vectors.

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

To highlight emotional information, we adopt NRC_VAD to compute emotion intensity values (Zhong, Wang, and Miao 2019b) for dialogue words and external concepts x:

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

where *min-max()* is min-max normalization; $\|.\|_k$ denotes L_k norm; $V_a(x)$ and $A_r(x)$ denote the values of valence and arousal dimensions in VAD vector of word x, respectively. If x is not in NRC_VAD, $\eta(x)$ will be set to 0.

We inject concepts with higher emotion intensity values from ConceptNet into KEMP to help emotion perception and expression.

Method

Overview

We provide a general overview of KEMP in Figure 4. KEMP consists of 3 phases: (A) emtional context graph, (B) emotional context encoder, and (C) emotion-dependency decoder. To summarize, we are given a dialogue history with M utterances, i.e., $\mathcal{D} = [X_1, \dots, X_M]$, as the input, where the i-th utterance $X_i = [x_0^i, \dots, x_{m_i}^i]$ is a sequence of m_i words. In phrase (A), we enrich the dialogue history $\mathcal{D}$ with external knowledge into an emotional context graph $\mathcal{G}$. In phrase (B), emotional signals e_p of $\mathcal{D}$ are distilled based on the embeddings and emotion intensity values from $\mathcal{G}$. Given e_p and $\mathcal{G}$, phrase (C) incorporates an emotional cross-attention mechanism to selectively learn the emotional dependencies. Subsequently, we generate an empathetic response $\mathcal{Y} = [y_1, \dots, y_n]$ with appropriate emotion and informative content.

Emotional context graph

We construct emotional context graph $\mathcal{G}$ by interacting with two-type external knowledge sources. Following Li et al. (2020), we flat dialogue history into a long word sequence and insert a `CLS` token at the start of the token sentence, i.e., $\mathcal{X} = [\texttt{CLS}, x_1, \dots, x_m]$. For each non-stopword word $x_i \in \mathcal{X}$, we first retrieve a set of candidate tuples $\mathrm{T}_i = \left\{ \tau_i^k = (x_i, r_i^k, c_i^k, s_i^k) \right\}_{k=1,\dots,K}$ from ConceptNet. Then we adopt three heuristic steps to refine the emotion-related knowledge: (1)We extract a subset $\hat{\mathrm{T}}_i \subset \mathrm{T}_i$ by filtering tuples with relevant relations for empathetic response (e.g., "Causes") and adequate confidence score (i.e., $s_i^k > 0.1$). (2)We rank tuples by the emotion intensity values $\{\eta(c_i^k)\}_{k=1,\dots,K}$ of retrieved concepts $\{c_i^k\}_{k=1,\dots,K}$. For each word x_i, we select top K' tuples as the emotional knowledge subgraph. (3)We apply 3 types of directed edges to connect vertices: (i) *temporary* edges between two successive words; (ii) *emotion* edges between a word x_i and its emotional concepts c_i^k; (iii) *globality* edges between `CLS` token and other vertices.

Finally, the dialogue history is enriched by emotional knowledge and represented as the emotional context graph $\mathcal{G}$. The words $x \in \mathcal{X}$ and the emotional concepts constitute the vertices $V = \{v_i\}_{i=1,\dots,e}$ of $\mathcal{G}$, where e is the number of vertices. The above edges among vertices are set to 1 in the adjacency matrix $\mathcal{A}$ of $\mathcal{G}$.

Emotional context encoder

Emotional context graph encoding. We first use a word embedding layer and a positional embedding layer (Vaswani et al. 2017) to convert each vertice $v_i \in \mathcal{G}$ 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 vertices in dialogue history or external knowledge is helpful. So we incorporate the vertice state embedding $\mathbf{E}_v(v_i)$ for vertice v_i. The vector representation of vertices v_i is the composition of three types of embeddings:

$$\mathbf{v}_i = \mathbf{E}_w(v_i) + \mathbf{E}_p(v_i) + \mathbf{E}_v(v_i). \tag{2}$$

Then we apply a multi-head graph-attention mechanism to update the vertice representations with emotional knowledge. Specifically, each vertice $\mathbf{v}_i$ is contextualized by attending to all its immediate neighbours $\{\mathbf{v}_j\}_{j \in \mathcal{A}_i}$:

$$\begin{aligned} \hat{\mathbf{v}}_i &= \mathbf{v}_i + \big\Vert_{n=1}^{H} \sum_{j \in \mathcal{A}_i} \alpha_{ij}^n \mathbf{W}_v^n \mathbf{v}_j, \\ \alpha_{ij}^n &= a^n(\mathbf{v}_i, \mathbf{v}_j), \end{aligned} \tag{3}$$

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

$$a^n(\mathbf{q}_i, \mathbf{k}_j) = \frac{\exp((\mathbf{W}_q^n \mathbf{q}_i)^\top \mathbf{W}_k^n \mathbf{k}_j)}{\sum_{z \in \mathcal{A}_i} \exp((\mathbf{W}_q^n \mathbf{q}_i)^\top \mathbf{W}_k^n \mathbf{k}_z)}, \tag{4}$$

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.

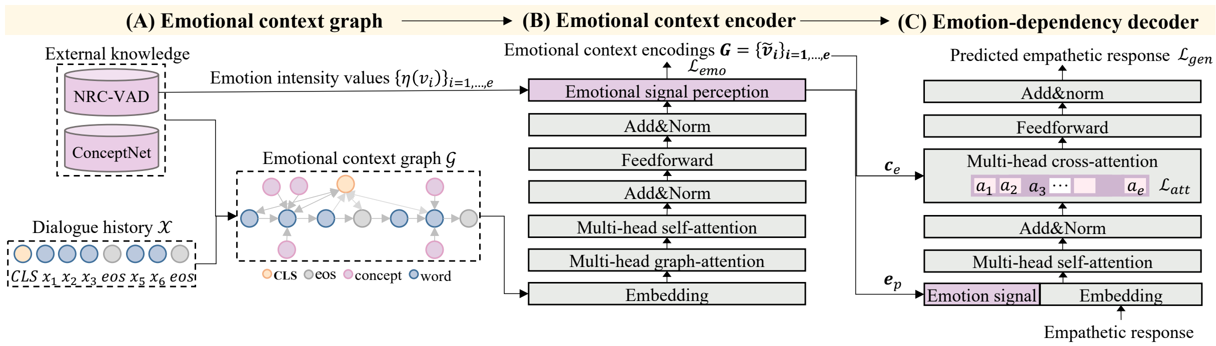

Figure 4: An overall architecture of KEMP. Model inputs are in the dotted box.

As previous operations are only conducted to the local context (i.e., immediate neighbours), we update the vertex representations with the global context information (i.e., all other vertices) to model global interactions. Concretely, we use transformer layers (Vaswani et al. 2017) to inject global information for all vertices $\{\hat{\mathbf{v}}_i\}_{i=1,\ldots,m}$:

$$\mathbf{h}_i^l = \text{LayerNorm}(\hat{\mathbf{v}}_i^{l-1} + \text{MHAtt}(\hat{\mathbf{v}}_i^{l-1})), \quad (5)$$

$$\hat{\mathbf{v}}_i^l = \text{LayerNorm}(\mathbf{h}_i^l + \text{FFN}(\mathbf{h}_i^l)), \quad (6)$$

where LayerNorm is the Layer Normalization trick (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 emotional context graph $\mathcal{G}$ is represented as $\mathbf{G} = \{\tilde{\mathbf{v}}_i\}_{i=1,\ldots,e}$, where $\tilde{\mathbf{v}}_i = \hat{\mathbf{v}}_i^l$.

Emotional signal perception. Our model learns the emotional signals from the emotional context graph to guide the empathetic response generation. The emotional signal representation $\mathbf{c}_e \in \mathbb{R}^d$ is the weighted summation of vertice representations $\{\tilde{\mathbf{v}}_i\}_{i=1,\ldots,e}$ on their emotion intensity values $\{\eta(v_i)\}_{i=1,\ldots,e}$:

$$\mathbf{c}_e = \sum_{i=1}^{m} \frac{\exp(\eta_i)}{\sum_{j=1}^{e} \exp(\eta_j)} \tilde{\mathbf{v}}_i. \quad (7)$$

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 to identify the emotional signal for the empathetic response:

$$\mathbf{e}_p = \mathbf{W}_e \mathbf{c}_e, \quad (8)$$

$$P_e(e|\mathcal{G}) = \text{softmax}(\mathbf{e}_p), \quad (9)$$

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

$$\mathcal{L}_{emo} = -\log(P_e(e = e^*|\mathcal{G})), \quad (10)$$

where e^* denotes the ground truth emotion label of dialogue history and e denotes the predicted label. Together with the emotional context encodings $\mathbf{G}$, emotional vectors $\mathbf{e}_p$ and $\mathbf{c}_e$ will be fed into the decoder as a crucial emotional signal to guild the empathetic response generation.

Emotion-dependency decoder

Starting from the intermediate emotional signal $\mathbf{e}_p \in \mathbb{R}^{1 \times q}$, we propose an emotion-dependency decoder to generate the target word sequentially. To acquire emotion dependencies from $\mathcal{G}$ and control empathetic response expression, we linearly transform $\mathbf{e}_p$ to $\mathbf{e}'_p$ via $\mathbf{e}'_p = \mathbf{W}_z \mathbf{e}_p + \mathbf{b}_z$. At the j-th decoding step, $\mathbf{e}'_p$ is concatenated with the embeddings of words $[y_1, \ldots, y_{j-1}]$ into $[\mathbf{y}_0, \ldots, \mathbf{y}_{j-1}]$, where $\mathbf{y}_0 = \mathbf{e}'_p$. We then feed the embeddings into the response decoder.

Our decoder is built based on Transformer layers. Specially, to improve the emotional dependencies between the emotional context graph and target empathetic response, we design two emotional strategies, i.e., *incorporating emotional features* and *enforcing emotional attention loss* at the cross-attention sub-layer.

Incorporating emotional features. To capture dialogue context vector $\mathbf{g}_s$ from emotional context graph $\mathcal{G}$, we compute the attention score between the last prediction word $\mathbf{y}_j$ and vertices $\{\tilde{\mathbf{v}}_i\}_{i=1,\ldots,e}$ as follows:

$$a^n(\mathbf{y}_{j-1}, \tilde{\mathbf{v}}_i) = \frac{\exp((\mathbf{W}_c^n \tilde{\mathbf{v}}_i)^\top \mathbf{W}_r^n \mathbf{y}_{j-1})}{\sum_{v_z \in \mathcal{G}} \exp((\mathbf{W}_c^n \tilde{\mathbf{v}}_z)^\top \mathbf{W}_r^n \mathbf{y}_{j-1})}, \quad (11)$$

$$\mathbf{g}_s = \Big\Vert_{n=1}^{H} a^n(\mathbf{y}_{j-1}, \tilde{\mathbf{v}}_i) \mathbf{W}_u^n \tilde{\mathbf{v}}_i, \quad (12)$$

where H is the number of attention heads. To improve the empathy expression of response, we concatenate the context vector $\mathbf{g}_s$ with the emotional signals $\mathbf{c}_e$ into an emotional context vector $\mathbf{c}$, i.e., $\mathbf{c} = [\mathbf{g}_s; \mathbf{c}_e]$.

Then we feed the last word representation $\mathbf{y}_{j-1}$ and vector $\mathbf{c}$ to a two-layer feed-forward network, which has a ReLU activation function and a highway layer normalization, so we have:

$$\mathbf{s}_{j-1} = \text{LayerNorm}(\mathbf{y}_{j-1} + \mathbf{c}), \quad (13)$$

$$\mathbf{y}_j = \text{LayerNorm}(\mathbf{s}_{j-1} + \text{FFN}(\mathbf{s}_{j-1})), \quad (14)$$

Enforcing emotional attention loss. Since humans naturally pay extra attention to the emotional salient information during a conversation (Li et al. 2020), we enforce an emotional attention loss to focus on those vertices with higher emotion intensity values:

$$a_i = \sum_{n}^{H} a^n(\mathbf{y}_{j-1}, \mathbf{v}_i)/H, \quad (15)$$

$$\mathcal{L}_{att} = \frac{1}{e}\sum_{i=1}^{e}(\eta(v_i) - a_i)^2, \quad (16)$$

Then the generator yields the distribution over the vocabulary $\mathcal{V}$ for the j-th word:

$$P_{\mathcal{V}}(y_j \mid \mathbf{y}_{0:j-1}, \mathbf{G}) = \text{softmax}(\mathbf{W}_v \mathbf{y}_j + \mathbf{b}_v), \quad (17)$$

where $\mathbf{W}_v \in \mathbb{R}^{|\mathcal{V}|\times d}$, $\mathbf{b}_v \in \mathbb{R}^{|\mathcal{V}|}$ are trainable parameters.

By using external concepts, we compute a probability p_g of copying from vertices $\{v_i\}_{i=1,...,e}$ in the graph $\mathcal{G}$ in a manner similar to See, Liu, and Manning (2017) and derive the final probability distribution $P(y_j)$:

$$p_{gen} = \sigma(\mathbf{W_g}\mathbf{y}_j + b_g), \quad (18)$$

$$P(y_j) = p_g P_{\mathcal{V}}(y_j) + (1 - p_g)\sum_{i:v_i=y_j} a_i, \quad (19)$$

where $\mathbf{W_g} \in \mathbb{R}^d$ and $b_g \in \mathbb{R}$ are trainable parameters; $\sigma(\cdot)$ is the sigmoid activation function. We use the negative log-likelihood of the ground-truth words y_j^* as the generation loss function:

$$\mathcal{L}_{gen} = -\sum_{j=1}^{n}\log P(y_j = y_j^* \mid y_{1,...,j-1}^*, \mathcal{G}). \quad (20)$$

Eventually, we adopt a multi-task learning framework to jointly minimize the emotion perception loss (Eq. 10), the emotional attention loss (Eq. 16), and the generation loss (Eq. 20) as follows:

$$\mathcal{L} = \gamma_1\mathcal{L}_{emo} + \gamma_2\mathcal{L}_{gen} + \gamma_3\mathcal{L}_{att}. \quad (21)$$

where $\gamma_1, \gamma_2, \gamma_3$ are hyper-parameters.

Experimental Settings

Dataset

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 dialogue history (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 dialogue history 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 lengths of dialogue history and response are 2.1 utterances and 13.5 tokens respectively.

Baselines for comparison

We compare with the state-of-the-art baselines as follows: (1) **Transformer** (Vaswani et al. 2017): A Transformer-based encoder-decoder model with a copy mechanism. (2) **EmoPrepend-1** (Rashkin et al. 2019): An extension of the Transformer model which incorporates an additional supervised emotion classifier. (3) **MoEL** (Lin et al. 2019): Another extension of Transformer model which softly combines the response representations from different decoders. Each decoder is optimized to focus on one type of emotion accordingly. (4) **MIME** (Majumder et al. 2020): An empathetic dialogue model considering polarity-based emotion clusters and emotional mimicry. (5) **EmpDG** (Li et al. 2020): A multi-resolution empathetic adversarial chatbot which exploits multi-resolution emotions and user feedback.

We also conduct ablation studies to better analyze the influence of different components in our model: (1) **w/o ECE**: The KEMP model without emotional knowledge of the emotional context encoder. (2) **w/o EDD**: The KEMP model without emotion-dependency mechanisms of the decoder. Additionally, we analyze the results of incorporating pre-trained model (DialoGPT (Zhang et al. 2020)) in our model.

Implementation details

We lowercase the characters, tokenize the sequences and retain a vocabulary with 24,647 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 (Li et al. 2020). The maximum introducing numbers of external concepts per dialogue and per token are set as 10 and 5, respectively. The threshold α used in emotional context graph construction is 0.1. Loss weights $\gamma_1, \gamma_2, \gamma_3$ are set to 1, 1, and 0.1, respectively. 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 2015) with a mini-batch size of 16. We varied the learning rate during training following Vaswani et al. (2017). Early stopping is applied when training. When inference, we set the maximum decoding step as 30. The training time of KEMP is 3 hours for around 26000 iterations.

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. Following previous emotion-related studies (Zhou et al. 2018a; Rashkin et al. 2019; Song et al. 2019; Wei et al. 2019; Li et al. 2020), we adopt **Perplexity** (Serban

Table 2: Performance of all models.

Models	Accuracy	Perplexity	Distinct-1	Distinct-2	Empathy	Relevance	Fluency
Transformer (Vaswani et al. 2017)	-	37.73	0.47	2.04	3.11	3.47	3.66
EmoPrepend-1 (Rashkin et al. 2019)	33.28	38.30	0.46	2.08	3.23	3.51	3.67
MoEL (Lin et al. 2019)	32.00	38.04	0.44	2.10	3.37	3.78	3.64
MIME (Majumder et al. 2020)	34.24	37.09	0.47	1.91	3.38	3.66	3.63
EmpDG (Li et al. 2020)	34.31	37.29	0.46	2.02	3.45	3.88	**3.67**
KEMP	**39.31**	**36.89**	**0.55**	**2.29**	**3.49**	**3.92**	3.65

Table 3: Ablation study.

Models	Accuracy	Perplexity	Distinct-1	Distinct-2
KEMP	**39.31**	36.89	**0.55**	**2.29**
w/o ECE	38.80	36.42	0.52	2.09
w/o EDD	35.41	**36.14**	0.41	2.04

Table 4: Result of human A/B test.

Models	Win	Loss	Tie
KEMP vs Transformer	43.8%	17.5%	38.7%
KEMP vs EmoP	40.6%	18.5%	40.9%
KEMP vs MoEL	38.3%	18.0%	43.7%
KEMP vs MIME	36.6%	20.6%	42.8%
KEMP vs EmpDG	35.5%	21.3%	43.2%

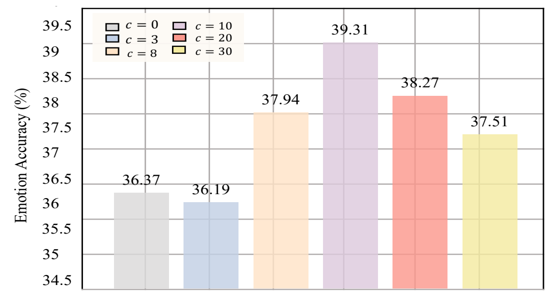

Figure 5: Emotion accuracy with respect to the maximum number of external concepts injection (c).

et al. 2015), **Distinct-1**, and **Distinct-2** (Li et al. 2016a) 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** (Lin et al. 2019; Majumder et al. 2020; Li et al. 2020). Empathy measures whether the generated responses express the appropriate emotions; Relevance evaluates whether the responses are on-topic with the dialogue history; 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.

Results and analysis

Automatic evaluation results

In Table 2, we observe that our model KEMP outperforms strong baselines MIME and EmpDG by a large margin in terms of all automatic metrics. The noticeable improvement indicates the effectiveness of our knowledge-enhanced model in empathetic expression and response diversity. EmpPrepend-1 and MoEL have similar performance, as both of them only use the dialogue history 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 replace emotional context encoder with vanilla transformer encoder (w/o ECE model), both the emotion accuracy and distinct performance become obviously worse, indicating that injecting external knowledge is consistently critical for emotion understanding and response generation. We also investigate the effect of replacing emotion-dependency decoder with vanilla transformer decoder (i.e., w/o EDD model). We notice that the scores decrease dramatically on most metrics, which demonstrates the effectiveness of modelling emotional dependencies.

Human evaluation results

Table 2 illustrates that KEMP obtains the best performance on both Empathy and Relevance scores. This suggests that the knowledge-enriched emotional context encoder and emotion-dependency decoder 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 KEMP are more preferred by human judges.

Table 5: The visualization of the cross-attention weights in EmpDG and KEMP.

History	It inspires me to try and do something to keep healthy every day .
EmpDG	I am sorry to hear. What kind of health is it?
History	It inspires me to try and do something to keep healthy every day .
Knowledge	effort , fight , good , life , raise , grow , protect , health
KEMP	I can not wait to <u>try</u> to get a little <u>makes</u> me <u>feel</u> <u>better</u>.

Table 6: Results on the pre-trained models.

Models	Accuracy	Perplexity	Distinct-1	Distinct-2
KEMP-big	45.91	-	2.22	4.93
DialoGPT	-	15.57	1.57	4.18
KEMP-DialoGPT	**46.43**	**15.21**	**2.79**	**4.24**

External knowledge analysis

To further investigate the impact of the different introduced number of external knowledge, we train KEMP with different numbers of concepts in terms of Accuracy. The result is 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 increases or even decreases. Therefore, external knowledge is more suitable to be the auxiliary information to perceive the emotional states in the dialogue history.

Emotion-dependency analysis

Table 5 shows an example illustrating the cross-attention weights of the dialogue context. Baseline EmpDG puts the major attention on general words, which leads to a context-inconsistent and emotion-inappropriate response. In comparison, the KEMP 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 dialogue history. We can conclude that the proposed emotion-dependency mechanism in the decoder can teach the model to generate responses from meaningful and emotional words.

Effectiveness of pre-trained model

As show in Table 6, we also explore if we can improve the performance by integrating KEMP with the pre-trained model on dialogues, i.e, DialoGPT (Zhang et al. 2020). KEMP-big is the KEMP with the same transformer hyperparameters setups as the DialoGPT. KEMP-DialoGPT incorporates the graph-attention layer of emotional context encoder and the cross-attention layer of emotion-dependency decoder into the DialoGPT. We can find that pre-trained models is effective in the empathetic dialogue genration because of the huge amount of pre-trained dialogue datasets. More important, we see that KEMP-DialoGPT outperforms DialoGPT, which concludes injecting emotional knowledge is able to improve the generation performance.

Table 7: Generated responses from KEMP and baseline models in two different speaker's emotion states. Tokens in <u>underline</u> represent knowledge-related words.

Emotion	**Terrified**
History	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.**
Transformer	I am sure it was.
EmoP	I am sure it was a lot of fun.
MoEL	I am sure you will do it!
MIME	That is a good thing to do.
EmpDG	I think it is an amazing feeling.
Know.	danger, pain, travel, <u>scared</u>, excited, <u>furious</u>
KEMP	I would have been so <u>scared</u>.
Emotion	**Surprised**
History	X_1: I realized yesterday I was pregnant, I was in shock!
Gold	**Woah, that is huge news. How are you handling the news.**
Transformer	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.
MIME	That is great.
EmpDG	Oh wow! what happened?
Know.	experience, <u>surprise</u>, mother, pain, <u>feeling</u>
KEMP	Oh wow! That is a <u>great</u> <u>feeling</u>.

Case study

Cases from KEMP and baseline models are listed in Table 7. In the first case, KEMP generates informative responses with a proper negative emotion by replying with "*scared*". However, without emotional knowledge, all baselines fail to recognize the negative emotion. In the second case, KEMP model generates the most context-consistent response, which contains context-related word ("*feeling*") and emotion-rated word ("*Oh wow*"). Both the two cases show that the KEMP can balance the performances between content and emotion.

Conclusion and outlook

In this work, we have proposed a knowledge-aware empathetic dialogue generation model, KEMP, to enhance the emotion perception and dependencies abilities of empathetic dialogue system with bunches of emotion-related concepts. Experimental results show that KEMP outperforms state-of-the-art methods in terms of both automatic and human evaluations. Besides, we verify the effectiveness of the emotional context graph, emotional context encoder, and the emotion-dependency decoder in KEMP.

KEMP adopts heuristic rules to construct emotional context graph, which is not flexible to adapt different knowledge resources. As for future work, we plan to address this issue by integrating with knowledge reasoning models to automatically construct emotional context graph.

Acknowledgments

We want to thank our anonymous reviewers for their feedback. This work was supported by the National Key R&D Program of China with grant No. 2020YFB1406704, the Natural Science Foundation of China (62106105, 62102234, 62072279, 61902219, 61972234), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), the Natural Science Foundation of Shandong Province (ZR2021QF129), the Fundamental Research Funds of Shandong University.

References

Asghar, N.; Poupart, P.; Hoey, J.; Jiang, X.; and Mou, L. 2018. Affective Neural Response Generation. In *ECIR*, 154–166.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *stat*, 1050: 21.

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.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *ICLR*.

Ghosal, D.; Hazarika, D.; Roy, A.; Majumder, N.; Mihalcea, R.; and Poria, S. 2020. KinGDOM: Knowledge-Guided DOMain Adaptation for Sentiment Analysis. In *ACL*, 3198–3210.

Hsu, C.-C.; Chen, S.-Y.; Kuo, C.-C.; Huang, T.-H.; and Ku, L.-W. 2018. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *LREC*.

Huang, C.; Zaiane, O. R.; 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. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, W. B. 2016a. 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. 2016b. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*, 1192–1202.

Li, J.; and Sun, X. 2018. A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation. In *EMNLP*, 678–683.

Li, Q.; Chen, H.; Ren, Z.; Ren, P.; Tu, Z.; and Chen, Z. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *COLING*, 4454–4466.

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*, 121–132.

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*, volume 34, 13622–13623.

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A. F.; Mihalcea, R.; and Poria, S. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *EMNLP*, 8968–8979.

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*, 469–474.

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. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

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.

Santhanam, S.; and Shaikh, S. 2019. Emotional Neural Language Generation Grounded in Situational Contexts. In *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, 22–27.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *ACL*, 1073–1083.

Serban, I. V.; Lowe, R.; Charlin, L.; and Pineau, J. 2016. Generative Deep Neural Networks for Dialogue: A Short Review. *CoRR*, abs/1611.06216.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8): 434–441.

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. In *ACL*, 556–566.

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*, 4444–4451.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 3104–3112.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NeurIPS*, 5998–6008.

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.

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.

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.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *ACL*, 270–278.

Zhong, P.; Wang, D.; and Miao, C. 2019a. An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss. In *AAAI*, 7492–7500.

Zhong, P.; Wang, D.; and Miao, C. 2019b. Knowledge-Enriched Transformer for Emotion Detection in Textual Conversations. In *EMNLP-IJCNLP*, 165–176.

Zhong, P.; Zhang, C.; Wang, H.; Liu, Y.; and Miao, C. 2020. Towards Persona-Based Empathetic Conversational Models. In *EMNLP*, 6556–6566.

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*, 730–739.

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.

Zhou, X.; and Wang, W. Y. 2018. MojiTalk: Generating Emotional Responses at Scale. In *ACL*, 1128–1137.