Constructing Emotion Consensus and Utilizing Unpaired Data for Empathetic Dialogue Generation

Lei Shen\textsuperscript{1,2}  Jinchao Zhang\textsuperscript{2}  Jiao Ou\textsuperscript{1,2}  Xiaofang Zhao\textsuperscript{1*}  Jie Zhou\textsuperscript{3}
\textsuperscript{1}Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
\textsuperscript{2}University of Chinese Academy of Sciences, Beijing, China
\textsuperscript{3}Pattern Recognition Center, WeChat AI, Tencent Inc, China
{shenlei17z, oujiao17b, zhaoxf}@ict.ac.cn, {dayerzhang, withtomzhou}@tencent.com

Abstract

Researches on dialogue empathy aim to endow an agent with the capacity of accurate understanding and proper responding for emotions. Existing models for empathetic dialogue generation focus on the emotion flow in one direction, that is, from the context to response. We argue that conducting an empathetic conversation is a bidirectional process, where empathy occurs when the emotions of two interlocutors could converge on the same point, i.e., reaching an emotion consensus. Besides, we also find that the empathetic dialogue corpus is extremely limited, which further restricts the model performance. To address the above issues, we propose a dual-generative model, Dual-Emp, to simultaneously construct the emotion consensus and utilize some external unpaired data. Specifically, our model integrates a forward dialogue model, a backward dialogue model, and a discrete latent variable representing the emotion consensus into a unified architecture. Then, to alleviate the constraint of paired data, we extract unpaired emotional data from open-domain conversations and employ Dual-Emp to produce pseudo paired empathetic samples, which is more efficient and low-cost than the human annotation. Automatic and human evaluations demonstrate that our method outperforms competitive baselines in producing coherent and empathetic responses.

1 Introduction

Empathy, a fundamental trait of humans, describes the ability to place oneself in another person’s position and share his/her feelings or emotions. Besides, it has been considered to be one of the most valuable affective phenomena for improving human-machine interactions (Zech and Rimé, 2005). The studies of empathy in natural language processing mainly include detecting empathy in spoken language or text (Buechel et al., 2018; Sharma et al., 2020), generating empathetic dialogue responses (Lin et al., 2019; Majumder et al., 2020), and constructing empathy lexicons (Sedoc et al., 2020) or datasets (Rashkin et al., 2019).

The empathetic dialogue generation task has been regarded as a unidirectional process from the context to response, and is modeled as a multi-task learning that combines the emotion understanding and the emotion-enhanced response generation. Therefore, existing work (Rashkin et al., 2019; Lin et al., 2019; Li et al., 2020; Majumder et al., 2020) mainly focuses on improving the accuracy of emotion classification or enhancing response generation via integrating the detected emotion factor.

Conducting an empathetic conversation is naturally a bidirectional process: the speaker conveys his/her emotion by describing a certain situation, then the listener receives that emotion and feeds his/her feeling back to the listener via a response. Then, the empathy is triggered when two interlocutors link similar experiences and their emotions could converge on the same point, i.e., reaching an emotion consensus. Take the case in Figure 1 as an example:

![Figure 1: An example of conducting an empathetic conversation. Both responses show empathy to the speaker.](image-url)
example. The emotion consensus “Sad” works as an intersection that connects both the speaker and listener, and it is a high-level abstraction behind the content, i.e., both two responses convey their acknowledgment of the sad feeling even with different expressions. Therefore, a unidirectional model is not enough to model the relationship between the context and response. Besides, previous models for this task only utilize paired data with limited capacity in a benchmark dataset, EMPATHETICDIALOGUES. Rather than manually annotating a larger empathetic dataset, we find that in open-domain conversations, there is large-scale emotional data that can be used to improve the performance. Compared with recognizing whether a context-response pair is empathetic, obtaining either an emotional context or response (named as unpaired data in this paper) can be easier with a well-trained classifier.

In this paper, we propose a Dual-Generative model for the Empathetic dialogue generation task (Dual-Emp), which simultaneously constructs emotion consensus and utilizes unpaired data. Dual-Emp combines a forward dialogue model (generating a response based on its context) and a backward dialogue model (generating a context based on its responses) with a discrete latent variable. Specifically, the forward and backward encoders convert the context and response into vectors at the same time, and then a discrete latent variable is used to capture the high-level emotion consensus shared in each context-response pair. Moreover, the latent variable and an emotion-enhanced attention mechanism are integrated into both forward and backward decoders to better express proper empathy. To utilize unpaired emotional data, we firstly extract them from open-domain conversations with emotions. Then we can get pseudo pairs by feeding either emotional responses or contexts to the backward or forward model. A joint training process is introduced to promote the semantic coherence between contexts and responses. Furthermore, two types of optimization methods are applied to better train the entire model with paired and unpaired data. Experimental results on a benchmark dataset EMPATHETICDIALOGUES show that Dual-Emp significantly outperforms competitive baselines in generating meaningful and related responses while expressing an appropriate empathy.

Our main contributions can be summarized as:

(1) We point out that the empathetic dialogue generation contains bidirectional processes, and highlight the importance of constructing emotion consensus. Besides, we propose a novel dual-generative model that couples a forward and a backward dialogue model with a discrete latent variable capturing the shared emotion consensus. (2) We utilize unpaired emotional data to break the constraint of paired empathetic data in the widely-used benchmark dataset EMPATHETICDIALOGUES. (3) Automatic and human evaluations show that our model outperforms competitive baselines in terms of fluency, coherence, and empathy.

2 Related Work

Emotion-Controllable Response Generation. Infusing emotions into dialogue systems can make conversational agents more human-like and benefit the interactions between human and machine (Prendinger and Ishizuka, 2005). Emotion-controllable response generation aims to generate emotional responses conditioning on a manually-provided label. Existing work (Zhou et al., 2018; Zhou and Wang, 2018; Colombo et al., 2019; Song et al., 2019; Shen and Feng, 2020) focused on obtaining responses that are not only meaningful, but also in accordance with the desired emotion.

Empathetic Response Generation. Rashkin et al. (2019) considered a richer and evenly distributed set of emotions, and released a dataset EMPATHETICDIALOGUES. Shin et al. (2020) formulated a reinforcement learning problem to maximize user’s sentimental feelings towards the generated responses. Lin et al. (2019) presented an encoder-decoder model with each emotion having a dedicated decoder. Majumder et al. (2020) introduced emotion grouping, emotion mimicry, and stochasticity to generate empathetic and various responses. Li et al. (2020) integrated knowledge to better understand dialogue contexts, and also designed an emotion-focused attention mechanism for emotional dependencies.

Dual Learning in NLP. He et al. (2016) proposed Dual Learning (DL) for machine translation first, which considered the source to target language translation and target to source language translation as a dual task. After that, Tang et al. (2017) implemented a dual framework for the question-answering system. Both Zhang et al. (2018a) and Cui et al. (2019) used similar idea in dialogue generation task to produce coherent but not safe responses. Shen and Feng (2020) applied DL for emotion-controllable response generation with
three awards for emotions and semantics. Some researchers also exploited DL to relieve the need of paired data and make use of unpaired data in several areas, such as style transfer (Luo et al., 2019a,b), semantic understanding (Tseng et al., 2020), stylized response generation (Zheng et al., 2020a), and machine translation (Zheng et al., 2020b).

The differences between our model and previous methods are: (1) To improve the empathy understanding, we introduce a backward model to represent the response and a discrete latent variable to capture the emotion consensus shared by contexts and responses. (2) Our forward and backward models are connected by a latent variable, and both of them can be updated at each iteration, while traditional DL can only fix one to update another.

3 Proposed Method

For empathetic dialogue generation, a dialogue consists of utterances from a speaker and a listener. Given context \( c = \{S_1, L_1, S_2, L_2, \ldots, S_t\} \), where \( S_i = \{w_j^{s_i}\}_{j=1}^{|S_i|} \) denotes speaker and \( L_i = \{w_j^{l_i}\}_{j=1}^{|L_i|} \) denotes listener, the goal is to track the speaker’s emotion state from \( c \), and generate a response \( y = L_t \) that is meaningful and empathetic.

3.1 Overview

The architecture of Dual-Emp is shown in Figure 2. Dual-Emp has five modules, the forward encoder \( f_{enc} \), forward decoder \( f_{dec} \), backward encoder \( b_{enc} \), backward decoder \( b_{dec} \), and \( z_e \) indicating a discrete latent variable. \( z_e \) can be inferred from both \( c \) and \( y \) and is used to capture emotion consensus shared in each \((c, y)\) pair. Because of the existence of \( z_e \), other modules are correlated and can better model both the semantic relation and the emotion connection between \( c \) and \( y \).

3.2 Model Architecture

Since the backward dialogue model has the same architecture as the forward one, we specify the components of forward dialogue model below and omit those of backward model for space limitation. **Encoder.** Following the work of Lin et al. (2019), we firstly concatenate utterances in \( c \) into a long sequence with length \( n \) and add a special token CTX to the beginning of \( c \) inspired by BERT (Devlin et al., 2019). Then, each token \( w \) in \( c \) is calculated as the sum of three embeddings:

\[
E_c(w) = E_w(w) + E_p(w) + E_r(w), \tag{1}
\]

where \( E_w(\cdot) \), \( E_p(\cdot) \), and \( E_r(\cdot) \in \mathbb{R}^{|V| \times d_{emb}} \) represent word embedding space, positional embedding space and role embedding space\(^1\), respectively. Finally, a transformer encoder (Vaswani et al., 2017) \( f_{enc} \) is applied to get the context representation:

\[
H = f_{enc}(E_c([CTX; c])), \tag{2}
\]

where \(";\) represents the concatenation operation, and \( H \in \mathbb{R}^{(n+1) \times d_{mod}} \). The contextualized encoding of CTX, i.e., \( H_0 \in \mathbb{R}^{d_{mod}} \), is used as the final representation of the entire context.

**Emotion Consensus Construction.** A \( K \)-way categorical latent variable \( z_e \in \{1, K\} \) (Bao et al., 2020) is used to capture the emotion consensus shared by \( c \) and \( y \). Inspired by Zhao et al. (2019), we define the prior distribution where we sample \( z_e \) from to be uniform\(^2\), i.e., \( p(z_e) = 1/K \). Correspondingly, the approximate posterior distribution is defined as follows:

\[
q(z_e|c) = \text{softmax}(\text{FFN}(H_0)) \in \mathbb{R}^K, \tag{3}
\]

where \( \text{FFN}() \) represents a feedforward network. This part can be considered as the emotion understanding on \( c \). Here \( z_e \) has its own embedding space \( E_z \in \mathbb{R}^{K \times d_{mod}} \) to convert it into a vector, i.e., \( E_z(z_e) \in \mathbb{R}^{d_{mod}} \). To supervise the emotion expression in \( E_z(z_e) \), we train a classifier using the cross-entropy loss between \( E_z(z_e) \) and ground-truth emotion label \( e^* \):

\[
p_e = \text{softmax}(W_e E_z(z_e)), \tag{4}
\]

\[
L_{emo} = -e^* \log p_e, \tag{5}
\]

where \( W_e \) is a trainable weight matrix.

---

\(^1\)The roles in \( c \) is an alternating set of “speaker” and “listener”, while in \( y \), the role is “listener” only.

\(^2\)Since emotion labels in EMPATHETIC_DIALOGUES are evenly-distributed, we set the prior distribution to be uniform.
We firstly describe how Dual-Emp can be trained with the paired data \((c, y)\), and also the unpaired data \(c\) or \(y\). Then a combined objective is derived to optimize Dual-Emp using the paired and unpaired data at the same time.

**Training with Paired Data.** Given \((c, y)\), we aim to maximize the log-likelihood of a joint probability \(p(c, y)\):

\[
\log p(c, y) = \log \sum_{y_e} p(c, y, z_e). \tag{10}
\]

Following the derivations from Zhao et al. (2018), Tseng et al. (2020), and the variational inference (Kingma and Welling, 2014), an objective based on the evidence lower bound can be derived as:

\[
\mathcal{L}_1 = \mathbb{E}_{q(z_e|c)} \log p(y|z_e, c) + \mathbb{E}_{q(z_e|c)} \log p(c|z_e, y) - D_{\text{KL}}[q(z_e|c)||p(z_e)], \tag{11}
\]

where the first term denotes the forward dialogue model, \(q(z_e|c)\) is the approximate posterior distribution of \(z_e\), and is computed by the forward encoding process (red ⊙ in Figure 3(c)). \(p(y|z_e, c)\) is the forward decoding process (green ⊙ in Figure 3(c)); the second term denotes the reconstruction of \(c\) and \(p(c|z_e, y)\) is the backward decoding process (blue ⊙ in Figure 3(c)); the third term is a Kullback-Leibler (KL) divergence between two distributions.

Analogously, the posterior distribution of \(z_e\) can be approximated by \(q(z_e|y)\), and the objective can be converted as follows:

\[
\mathcal{L}_2 = \mathbb{E}_{q(z_e|y)} \log p(c|z_e, y) + \mathbb{E}_{q(z_e|y)} \log p(y|z_e, c) - D_{\text{KL}}[q(z_e|y)||p(z_e)], \tag{12}
\]

where terms have similar meanings to those in Eq. 11, and we only need to interchange “forward” and “backward”. Besides, the forward encoding process (red ⊙ in Figure 3(c)) is replaced with the backward encoding process (gray ⊙ in Figure 3). Detailed derivations can be found in Appendix. Therefore,
the final loss function for the paired data is:

\[ L_{cy} = L_1 + L_2 + \alpha L_{emo}. \]  

(13)

where \( \alpha \) is a hyper-parameter.

**Training with Unpaired Data.** Given unpaired data \( c \) (\( c \) is an emotional context), we need to maximize the log-likelihood of a marginal probability \( p(c) \):

\[ \log p(c) = \log \int_y \sum_{z_e} p(c, y, z_e). \]  

(14)

Then, we can get the evidence lower bound for the marginal probability:

\[ L_3 = \mathbb{E}_{q(y|z_e,c)} \mathbb{E}_{q(z_e|c)} \log p(c|z_e, y) - D_{KL}[q(z_e|c) || p(z_e)], \]  

where the first term is the reconstruction of \( c \), \( q(z_e|c) \) is computed by the forward encoding process (\( \circ \) in Figure 3(d)), \( q(y|z_e, c) \) is the forward generation process (\( \triangledown \) in Figure 3(d)), and \( p(c|z_e, y) \) is the backward decoding process (\( \triangledown \) in Figure 3(d)); the second term is a KL divergence.

The forward generation process \( q(y|z_e, c) \) is similar to the back-translation in machine translation (Zhang et al., 2018b), and we use \( f_{\text{dec}} \) to generate pseudo \( y' \) given \( c \) and \( z_e \). Since the ground-truth \( y \) is unobserved here, we apply reinforcement learning and policy gradient method (Williams, 1992) for training. The reward is designed as the probability of the model to reconstruct \( c \) based on the generated \( y' \) and \( z_e \):

\[ r = p(c|y', z_e). \]  

(16)

Similarly, we can get an objective when utilizing unpaired \( y \) (the emotional response):

\[ L_4 = \mathbb{E}_{q(e|z_e,y)} \mathbb{E}_{q(z_e|y)} \log p(y|z_e, c) - D_{KL}[q(z_e|y) || p(z_e)], \]  

(17)

where the first term is the reconstruction of \( y \), and the process is symmetrical to that of \( L_3 \). Detailed derivations can be found in Appendix. The final loss functions for unpaired data \( c \) and \( y \) are:

\[ L_c = L_3 + \beta L_{emo}, \]  

(18)

\[ L_y = L_4 + \gamma L_{emo}, \]  

(19)

where \( \beta \) and \( \gamma \) are two hyper-parameters.

**Total Training Loss.** During training, the paired empathetic data in EMPATHETICDIALOGUES and the unpaired emotional data from open-domain conversations are used simultaneously. Then, the total loss can be summarized as:

\[ L = L_{cy} + L_c + L_y. \]  

(20)

**Inference.** During inference, given the input \( c \), only the forward dialogue model is applied. We use \( f_{\text{enc}} \) to encode \( c \) and infer \( z_e \), then employ \( f_{\text{dec}} \) to generate \( y' \) based on \( c \) and \( z_e \).

### 4 Experiments

In this section, we conduct experiments to evaluate our proposed method. We firstly introduce some empirical settings. Then we illustrate our results on both automatic and human evaluations. Finally, we show some cases generated by different models and do further analyses over our method.

#### 4.1 Dataset

We conduct our experiments on the EMPATHETICDIALOGUES (Rashkin et al., 2019) dataset that consists of 24,850 conversations between two interlocutors. Each conversation in the dataset contains one emotion label, a situation where the speaker feels the exact emotion, and utterances about the speaker’s descriptions of the situation or the listener’s empathetic replies. There are 32 evenly-distributed emotion labels in the dataset. We apply the data provided by the original paper with the split ratio of 8:1:1 for training/validation/test set, and use the script released by Lin et al. (2019) to preprocess the data. Emotion labels are given as supervised signals in the training process, while during inference, they are predicted to evaluate the accuracy of emotion understanding.

#### 4.2 Implementation Details

We optimize the models using Adam (Kingma and Ba, 2015) with a mini-batch size of 16. The learning rate is initialized to 1e-4 and we vary the learning rate following Vaswani et al. (2017). Similar to Lin et al. (2019), Li et al. (2020), and Majumder et al. (2020), we use pre-trained GloVe vectors (Pennington et al., 2014) to initialize the word embeddings. Besides, all common hyper-parameters are set the same as previous work, e.g., the hidden size \( d_{\text{mod}} \) and embedding size \( d_{\text{emb}} \) are set to 300. In order to alleviate the degeneration problem of variational framework, we apply KL annealing (Bowman et al., 2016) that is the same as in Zhou
We compare our approach with five representative baselines: (1) Multi-TRS (Rashkin et al., 2019): A transformer-based model trained with emotion classification loss in addition to MLE loss, and the emotion label is classified from the encoder output; (2) MoEL (Lin et al., 2019): An extension to Multi-TRS, which softly combines the output states of the appropriate decoders and generates an empathetic response. Each decoder is optimized to focus on a specific emotion; (3) EmpDG (Li et al., 2020): A model that exploits coarse- and fine-grained emotions and introduces an interactive adversarial learning framework to use user feedbacks; (4) DualVAE (Tran and Nguyen, 2018): A model with two decoders: one is for CVAE, and the other is for response auto-encoding; (5) MIME (Majumder et al., 2020): A model that integrates emotion grouping, emotion mimicry, and stochasticity strategies to generate varied responses. MIME is also the state-of-the-art model for empathetic response generation. To make fair comparisons, we do not apply methods based on pre-trained models here, as both Dual-Emp and the above mentioned ones are not based on pre-trained models. Note that model (1) to (5) can only utilize the paired data.

Additionally, we also design following models for ablation study: (6) Sing-Emp-Paired: A variation of Dual-Emp with only the forward model and paired empathetic data; (7) Dual-Emp-Paired: Dual-Emp with only paired empathetic data.

### 4.4 Evaluation Measures

**Automatic Metrics.** For automatic evaluation, we use followings metrics: (1) BLEU (Papineni et al., 2002); (2) Embedding-based scores (Average, Greedy, and Extrema)4 (Liu et al., 2016; Serban et al., 2017); (3) Perplexity (PPL) (Vinyals and Le, 2015); (4) Dist-1/2 (Li et al., 2016); (5) Emotion accuracy (the agreement between the ground-truth emotion labels and the predicted ones from Eq. 5). Emotion accuracy can be used to measure the ability of emotion understanding.

**Human Evaluation.** Firstly, we randomly sample 100 contexts and their corresponding responses from our model as well as the baselines. Next, we send pairs of the context and generated response from different models to three professional annotators without order. Annotators are asked to evaluate each pair independently based on three distinct metrics: Empathy, Relevance, and Fluency (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020). Empathy measures the degree of emotional understanding of context shown by the response; Relevance evaluates whether the generated responses are relevant on topic with the context; Fluency assesses the grammatical correctness and readability of the generated responses. Each metric is rated on five-scale with “5” represents the best performance.

**Human A/B Test.** In this part, we try to directly compare Dual-Emp with other baselines. We randomly sample 100 dialogues each for Dual-Emp vs. {Multi-TRS, MoEL, EmpDG, DualVAE, MIME}. Three annotators are given generated responses from either Dual-Emp or {Multi-TRS, MoEL, EmpDG, DualVAE, MIME} in random order, and are asked to choose the better response. They can either choose one of the responses or select “Tie” when the provided options are either both good or both bad. The result of each sample is determined by majority voting. Finally, we calculate the percentage of samples where the first or second model generates the better response and where these two models perform similarly.

---

4We employ a popular NLG evaluation project available at https://github.com/Maluuba/nlg-eval.
Table 1: Automatic and human evaluation results. The metrics BLEU, Average, Greedy, Extrema, Dist-1, Dist-2, Emotion accuracy, Empathy, Relevance, and Fluency are abbreviated as B, Avg, Gre, Ext, D1, D2, EA, Emp, Rel, and Flu, respectively. Results show that Dual-Emp achieves the best performance on all metrics, especially a large improvement in Dist-1/2, Emotion accuracy, and Empathy.

<table>
<thead>
<tr>
<th>Method</th>
<th>B</th>
<th>Avg</th>
<th>Gre</th>
<th>Ext</th>
<th>PPL</th>
<th>D1 (%)</th>
<th>D2 (%)</th>
<th>EA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-TRS</td>
<td>2.56</td>
<td>0.938</td>
<td>0.786</td>
<td>0.541</td>
<td>33.82</td>
<td>0.68</td>
<td>2.62</td>
<td>35.17</td>
</tr>
<tr>
<td>MoEL</td>
<td>2.80</td>
<td>0.945</td>
<td>0.793</td>
<td>0.537</td>
<td>37.81</td>
<td>0.56</td>
<td>2.70</td>
<td>35.38</td>
</tr>
<tr>
<td>Emp-DG</td>
<td>2.79</td>
<td>0.935</td>
<td>0.788</td>
<td>0.532</td>
<td>34.31</td>
<td>0.47</td>
<td>2.10</td>
<td>34.35</td>
</tr>
<tr>
<td>DualVAE</td>
<td>2.76</td>
<td>0.941</td>
<td>0.791</td>
<td>0.540</td>
<td>33.46</td>
<td>0.77</td>
<td>3.21</td>
<td>35.36</td>
</tr>
<tr>
<td>MIME</td>
<td>2.82</td>
<td>0.946</td>
<td>0.794</td>
<td>0.536</td>
<td>37.53</td>
<td>0.51</td>
<td>2.68</td>
<td>34.88</td>
</tr>
<tr>
<td>Sing-Emp-Paired</td>
<td>2.77</td>
<td>0.944</td>
<td>0.790</td>
<td>0.533</td>
<td>32.71</td>
<td>0.75</td>
<td>2.91</td>
<td>28.75</td>
</tr>
<tr>
<td>Dual-Emp-Paired</td>
<td>2.86</td>
<td>0.950</td>
<td>0.792</td>
<td>0.542</td>
<td>32.56</td>
<td>0.80</td>
<td>3.09</td>
<td>36.82</td>
</tr>
<tr>
<td>Dual-Emp</td>
<td>2.91</td>
<td>0.957</td>
<td>0.796</td>
<td>0.545</td>
<td>31.01</td>
<td>1.08</td>
<td>3.23</td>
<td>37.53</td>
</tr>
</tbody>
</table>

Table 2: Results of human A/B test. Pairwise comparisons show that responses from Dual-Emp are more preferred by humans than those from baselines.

<table>
<thead>
<tr>
<th>Dual-Emp vs.</th>
<th>Win</th>
<th>Loss</th>
<th>Tie</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-TRS</td>
<td>43%</td>
<td>27%</td>
<td>30%</td>
<td>0.563</td>
</tr>
<tr>
<td>MoEL</td>
<td>37%</td>
<td>32%</td>
<td>31%</td>
<td>0.548</td>
</tr>
<tr>
<td>Emp-DG</td>
<td>40%</td>
<td>28%</td>
<td>32%</td>
<td>0.506</td>
</tr>
<tr>
<td>DualVAE</td>
<td>39%</td>
<td>30%</td>
<td>31%</td>
<td>0.527</td>
</tr>
<tr>
<td>MIME</td>
<td>36%</td>
<td>32%</td>
<td>32%</td>
<td>0.569</td>
</tr>
</tbody>
</table>

4.5 Experimental Results
Automatic Evaluation Results. The automatic evaluation results are shown in the left part of Table 1. The top part is the results of all baseline models, and we can see that Dual-Emp outperforms other methods on all metrics (t-test, p-value < 0.05). The improvements of Dual-Emp on PPL, Dist-1/2, and Emotion accuracy are significant, indicating that it can improve emotion understanding, and also enhance content fluency and diversity simultaneously. MoEL, Emp-DG, and MIME have similar performance, as they try to either improve the emotion understanding or intensify the emotion-based response generation.

The bottom part of Table 1 shows the results of our ablation study. Comparisons between Sing-Emp-Paired and Dual-Emp-Paired show the effectiveness of capturing emotion consensus with the assistance of both backward model and discrete latent variable. Especially, the noticeable improvement of Emotion accuracy indicates the discrete latent variable used for emotion prediction can help better model the emotion consensus by taking contexts and responses into consideration. In addition, we can find that with the support of unpaired emotional data, Dual-Emp achieves better results than Dual-Emp-Paired.

Human Evaluation Results. Human evaluation in Table 1 illustrates that Dual-Emp obtains the best performance (t-test, p-value < 0.05) on all scores. This suggests that our bidirectional model with latent variable helps construct emotion consensus shared by contexts and responses, thus improving the topic consistency and evoking more empathetic expressions. Besides, as more unpaired emotional data is utilized, Dual-Emp can achieve better Fluency. Additionally, we carry out pairwise comparisons to directly compare the response quality in Table 2. The results confirm that responses from Dual-Emp are more preferred by humans. Agreements to measure the consistency among three annotators are calculated with Fleiss’ kappa (Fleiss and Cohen, 1973), and the kappa values indicate “moderate agreement” in our cases.

4.6 Case Study
Table 3 shows two examples generated by Dual-Emp and other baselines. In the first case, Dual-Emp generates the most context-consistent response with a proper “apprehensive” emotion by replying with words “scary” and “what happened”, whereas baselines fail to understand the negative emotion or express inappropriate contents. In the second case, Dual-Emp generates a coherent and informative response, which corresponds to a subtle emotion change of the context from “lost a job” to “hoping he can find a full time job soon”. The response is not only emotion-related, but also contains the correct personal pronoun “he” and keyword “job”.

4.7 Further Analysis
Effects of Backward Model and $z_e$. To gain an insight into the effectiveness of backward dialogue model and the latent variable $z_e$, we plot the Emotion accuracy score of each emotion label based on Sing-Emp-Paired and Dual-Emp-Paired in Figure 4.
<table>
<thead>
<tr>
<th>Emotion</th>
<th>Apprehensive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context</td>
<td>I went skydiving last summer with my partner. It was so scary!</td>
</tr>
<tr>
<td>Ground-truth Response</td>
<td>Wow, that is scary. Were you nervous?</td>
</tr>
</tbody>
</table>

| MoEL        | I am so sorry to hear that. I am glad you are okay! |
| Emp-DG      | Did you go check it out? |
| MIME        | Oh no! I am sorry to hear that. I hope you get it! |
| Dual-Emp    | That is scary! What happened to you! |

Table 3: Generated responses from MoEL, Emp-DG, MIME, and Dual-Emp. In Case 1, Dual-Emp generates the most context-consistent response with a proper “apprehensive” emotion by replying with words “scary” and “what happened”. In Case 2, Dual-Emp captures a subtle emotion change of the context from “lost a job” to “hoping he can find a full time job soon”. Besides, it contains the correct personal pronoun “he” and keyword “job”.

Figure 4: Emotion accuracy over 32 emotions of Sing-Emp-Paired and Dual-Emp-Paired. The accuracy of Sing-Emp-Paired is unbalanced among all emotions, while Dual-Emp-Paired can not only improve the overall accuracy, but also exhibit a relatively even performance.

Table 4: Automatic evaluation results based on the number of unpaired data with different s values. Results show that more unpaired data does not lead to better results as some labels are not adequate with a low confidence.

<table>
<thead>
<tr>
<th>s</th>
<th>#context</th>
<th>#response</th>
<th>B</th>
<th>Avg</th>
<th>Gre</th>
<th>Ext</th>
<th>PPL</th>
<th>D1 (%)</th>
<th>D2 (%)</th>
<th>EA (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>324,243</td>
<td>314,070</td>
<td>2.25</td>
<td>0.937</td>
<td>0.791</td>
<td>0.549</td>
<td>31.22</td>
<td>0.90</td>
<td>2.90</td>
<td>35.62</td>
</tr>
<tr>
<td>0.55</td>
<td>224,324</td>
<td>216,839</td>
<td>2.69</td>
<td>0.933</td>
<td>0.784</td>
<td>0.537</td>
<td>32.63</td>
<td>1.87</td>
<td>4.32</td>
<td>36.06</td>
</tr>
<tr>
<td>0.60</td>
<td>155,059</td>
<td>149,672</td>
<td>2.91</td>
<td>0.957</td>
<td>0.795</td>
<td>0.545</td>
<td>31.01</td>
<td>1.08</td>
<td>3.23</td>
<td>37.53</td>
</tr>
<tr>
<td>0.65</td>
<td>107,189</td>
<td>103,192</td>
<td>2.25</td>
<td>0.934</td>
<td>0.783</td>
<td>0.539</td>
<td>32.61</td>
<td>1.70</td>
<td>4.80</td>
<td>35.66</td>
</tr>
<tr>
<td>0.70</td>
<td>73,132</td>
<td>70,200</td>
<td>2.60</td>
<td>0.938</td>
<td>0.791</td>
<td>0.540</td>
<td>31.66</td>
<td>0.70</td>
<td>2.40</td>
<td>37.51</td>
</tr>
</tbody>
</table>

As we can see, for Sing-Emp-Paired, some emotion categories can achieve pretty high accuracy, but in general, the accuracy is unbalanced among all emotions, which indicates that z_e cannot construct the emotion consensus well by only considering the contexts. In contrast, Dual-Emp-Paired not only improves the overall Emotion accuracy, but also exhibits a relatively even performance over all 32 emotions. Therefore, z_e can better understand the emotion via capturing emotion consensus with both forward and backward dialogue models.

**Choices of the Unpaired Data.** The threshold s we use in previous experiments equals to 0.60. Here, we choose different options to show their influence on the empathetic dialogue generation task.

Table 4 shows that more unpaired emotional data does not lead to better results as some labels are not adequate with a low confidence. The emotion classifier we applied to label the utterances from Reddit and Twitter is based on a 32-category classification task, thus it is hard to get very accurate results. Though the predicted emotion labels are noisy, these samples are good enough to train our model in practice.

5 Conclusion and Future Work

In this paper, we propose a dual-generative model, Dual-Emp, to generate the empathetic response given a context. We point out that conducting an
empathetic conversation is a bidirectional process, and empathy is mainly reflected by emotion consensus between the context and response. Then we couple forward and backward dialogue models with a discrete latent variable denoting the emotion consensus. Moreover, we integrate unpaired emotional data from open-domain conversations into Dual-Emp to relieve the need of paired data. Experimental results on a benchmark dataset show that Dual-Emp can generate fluent, related, informative, and empathetic responses. As the future work, we will prove the effectiveness of our method based on pre-trained models, and analyze how classification errors in unpaired data affect the generation.

Acknowledgements

We would like to thank all the reviewers for their insightful and valuable comments and suggestions.

References


