# EmpHi: Generating Empathetic Responses with Human-like Intents

Anonymous ACL submission


## Abstract

In empathetic conversations, humans express their empathy to others with empathetic intents. However, most existing empathetic conversational methods suffer from a lack of empathetic intents, which leads to monotonous empathy. To address the bias of the empathetic intents distribution between empathetic dialogue models and humans, we propose a novel model to generate **emp**athetic responses with **h**uman-consistent empathetic **i**ntents, **EmpHi** for short. Precisely, EmpHi learns the distribution of potential empathetic intents with a discrete latent variable, then combines both implicit and explicit intent representation to generate responses with various empathetic intents. Experiments show that EmpHi outperforms state-of-the-art models in terms of empathy, relevance, and diversity on both automatic and human evaluation. Moreover, the case studies demonstrate the high interpretability and outstanding performance of our model.


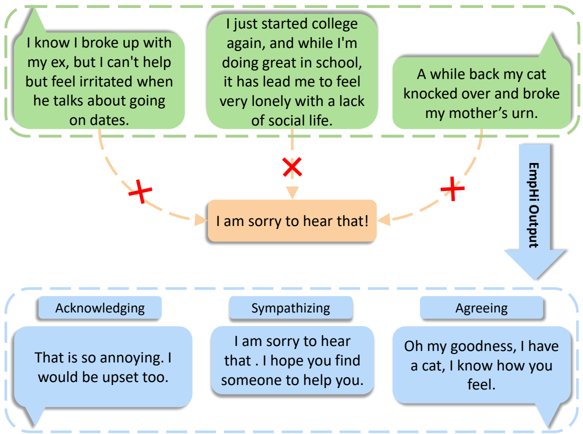


Figure 1: EmpHi generates empathetic responses with human-like empathetic intents (text in blue box), while existing empathetic dialogue models generate responses with contextually irrelevant and monotonous empathy (text in orange box).

## 1 Introduction

Empathy is a basic yet essential human ability in our daily life. It is a capacity to show one's caring and understanding to others. Many types of research have been conducted on empathetic expression to enhance the empathy ability of Artificial Intelligence, e.g., computational approach for empathy measurement (Sharma et al., 2020), empathetic expression understanding in newswire (Buechel et al., 2018), online mental health support (Sharma et al., 2021), etc. In this work, we focus on the task of generating empathetic responses (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020) in open-domain conversation.

Existing empathetic dialogue models pay more attention to the emotion-dependent response generation (Lin et al., 2019; Majumder et al., 2020). However, using emotion alone to generate responses is coarse-grained, and the model needs to incorporate empathetic intent information. On the one hand, the speaker often talks with a particular emotion while the listener shows their empathy with specific empathetic intents, e.g., *Acknowledging*, *Agreeing*, *Consoling* and *Questioning* etc (Welivita and Pu, 2020). On the other hand, see in Figure 1, when the user expresses sadness, existing models tend to generate sympathetic responses like "I'm sorry to hear that." However, empathy is not the same as sympathy, so the models should not only generate responses with *Sympathizing* intent. We demonstrate this phenomenon elaborately with a quantitative evaluation in Section 2. In real life situation, humans could reply with various empathetic intents to the same context which depends on personal preference. For example, given a context, "I just failed my exam", an individual may respond "Oh no, what happened?" with *Questioning* intent to explore the experience of the user, or "I understand this feeling, know how you feel" with *Agreeing* intent. These types of empathy are more relevant, interactive, and diverse.

To address the above issues, we propose a new framework to generate empathetic responses with human-like empathetic intents (EmpHi) which could generate responses with various empathetic intents, see examples in Figure 1. Specifically, EmpHi learns the empathetic intent distribution with a discrete latent variable and adopts intent representation learning in the training stage. During the generation process, EmpHi first predicts a potential empathetic intent and then combines both implicit and explicit intent representation to generate a response corresponding to the predicted intent. Our main contributions are:

- We discover and quantify the severe bias of empathetic intents between existing empathetic dialogue models and humans. This finding will lead subsequent researchers to pay more attention to fine-grained empathetic intents.
- To address the above problem, we propose EmpHi, which generates responses with human-like empathetic intents. Experiments have proved the effectiveness of our model through the significant improvement in both automatic and human evaluation.
- According to the quantitative evaluation and analysis, EmpHi successfully captures humans' empathetic intent distribution, and shows high interpretability in generation process.

## 2 Related Work

**Empathetic Response Generation.** Providing dialogue agents the ability to recognize speaker feelings and reply according to the context is challenging and meaningful. Rashkin et al. (2019) propose the **EmpatheticDialogues** for empathetic response generation research. Most subsequent empathetic conversation researches are evaluated on this dataset, including ours. They also propose Multitask-Transformer, which is jointly trained with context emotion classification and response generation. To further capture the fine-grained emotion information, Lin et al. (2019) introduce MoEL, a transformer with a multi-decoder. Each of them is responsible for the response generation of one specific emotion. Majumder et al. (2020) propose MIME, which mimics the speaker emotion to a varying degree.

All these models focus on emotion-dependent empathetic response generation, whereas we pay more attention to the empathetic intents and propose to generate a response that is not only emotionally appropriate but also empathetically human-like.

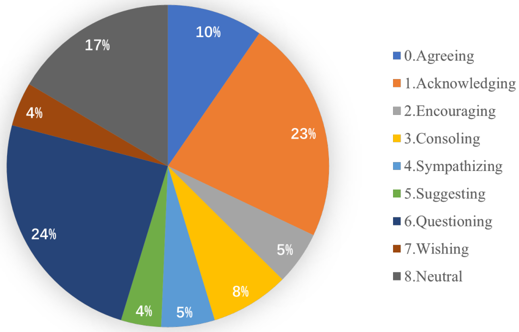


Figure 2: Empathetic intent distribution of human in empathetic conversation.

**One-to-many Response Generation.** Given dialogue history, there could be various responses depends on personal preference. Zhao et al. (2017) propose to learn the potential responses with continuous latent variable and maximize the log-likelihood using Stochastic Gradient Variational Bayes (SGVB) (Kingma and Welling, 2014). To further improve the interpretability of response generation, Zhao et al. (2018) propose to capture potential sentence-level representations with discrete latent variables. MIME (Majumder et al., 2020) introduces stochasticity into the emotion mixture for various empathetic responses generation.

Different from the previous works, we propose a discrete latent variable to control the empathetic intent of response and achieve intent-level diversity.

## 3 Empathetic Expression Bias

Although existing empathetic conversational methods have shown promising progress, we reveal there is a severe bias of empathetic expression between them and humans according to quantitative evaluation.

Empathy plays a vital role in human conversation, Welivita and Pu (2020) make a taxonomy of empathetic intents and calculate the frequency of each intent in **EmpatheticDialogues** (Rashkin et al., 2019). As shown in Figure 2, humans show their empathy naturally by *Questioning*, *Acknowledging*, and *Agreeing* intents etc.

However, there are no empirical experiments have shown *how empathetic dialogue models ex-*

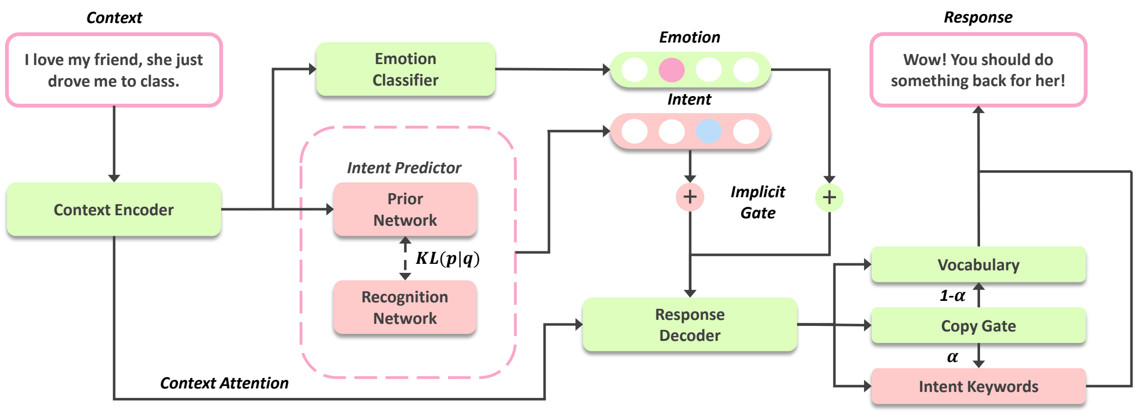


Figure 3: The architecture of EmpHi, which consists of a context encoder, an emotion classifier, a prior network (intent predictor), a recognition network, and a response decoder with copy mechanism.

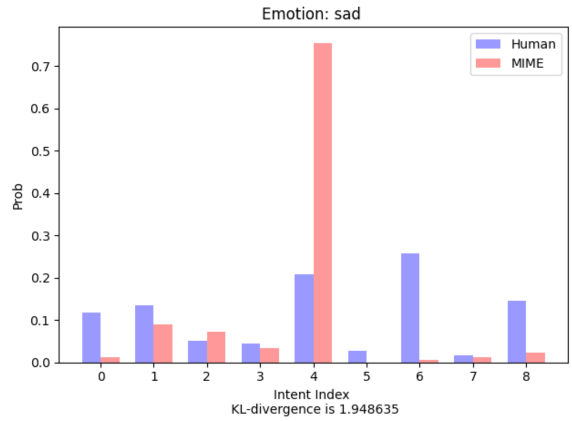


Figure 4: Empathetic intent distribution of human and MIME (sad emotion), the intent index represents the same intent as in Figure 2.

*press their empathy?* To further study, we finetune a BERT classifier (Devlin et al., 2019) on the released **EmpatheticIntents**[1] dataset (Welivita and Pu, 2020). Our classifier achieves 87.75% accuracy in intent classification and we apply it to identify the empathetic intents of responses generated by the state-of-the-art empathetic dialogue model MIME (Majumder et al., 2020). As shown in Figure 4, the severe empathetic intent distribution bias emerges while comparing humans to MIME. Given context with sad emotion, existing models usually generate "I am sorry to hear that" with *Sympathizing* intent, which is not human-like and contextually relevant. In addition, we can tell that the empathetic expression of MIME is monotonous. We also quantify the intent distribution of other empathetic dialogue models in the Appendix. The results are similar to Figure 4.

We believe this phenomenon is caused by that existing models only generate responses according to context emotion and lack fine-grained empathetic intent modeling. Therefore, we propose EmpHi, which generates empathetic responses with human-like empathetic intents.

[1] https://github.com/anuradha1992/EmpatheticIntents

## 4 EmpHi Method

### 4.1 Task Definition and Overview

Given the context, $C = [c_1, c_2, \cdots, c_m]$, which consists of $m$ words for single or multiple utterances. We aim to generate empathetic response, $X = [x_1, x_2, \cdots, x_n]$, with human-like empathetic intent. The whole model architecture is shown in Figure 3.

EmpHi learns the potential empathetic intent distribution with a latent variable $z$, which could be seen in Figure 5. Conditional Variational AutoEncoder (CVAE) (Yan et al., 2016; Zhao et al., 2017; Gu et al., 2019) is trained to maximize the conditional log likelihood, $\log p(X|C)$, which involves an intractable marginalization over $z$. We train the CVAE efficiently with *Stochastic Gradient Variational Bayes* (SGVB) (Kingma and Welling, 2014) by maximizing the variational lower bound of the log likelihood:

$$\begin{aligned}\log p(X|C) \geq &\mathbf{E}_{q(z|X,C)}[\log p(X|C,z)] \\ &- \mathbf{KL}(q(z|X,C)||p(z|C)),\end{aligned} \tag{1}$$

$p(X|C,z)$ denotes response reconstruction probability, $q(z|X,C)$ is recognition probability and $p(z|C)$ is prior probability. Our method mainly consists of three aspects:

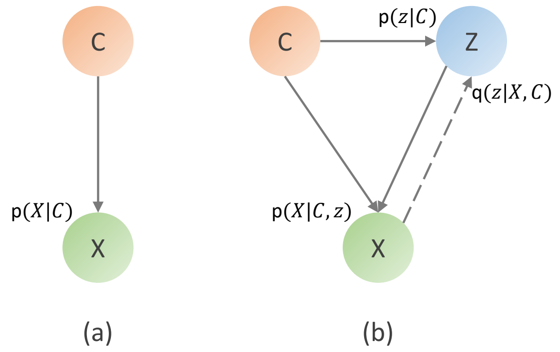


Figure 5: An illustration of the difference between existing empathetic dialogue models (a) and EmpHi (b).

- To capture the explicit relationship between the latent variable and the intent, we propose an intent representation learning approach to learn the intent embeddings.
- We construct an intent predictor to predict potential response intent using contextual information and then use this intent for guiding the response generation.
- During the generation process, EmpHi combines both implicit intent embedding and explicit intent keywords to generate responses corresponding to the given intents.

### 4.2 Learning Intent Representation

To achieve more interpretability, we choose a discrete latent variable that obeys categorical distribution with nine categories, each corresponding to one empathetic intent. Directly maximizing Eq.1 would cause two serious problems: the relation between the latent variable and intent is intractable; the *vanishing latent problem* results in insufficient information provided by the latent variable during generation. (Bowman et al., 2016; Zhao et al., 2017; Gu et al., 2019).

To solve the above issues, we separately train a recognition network $q_r(z|X)$ to encourage intent variable $z$ to capture context-independent semantics, which is essential for $z$ to be *interpretable* (Zhao et al., 2018). The task of the recognition network is to provide the accurate intent label of the response, which corresponds to an intent embedding. Then, by maximizing likelihood $p(X|C, z)$, the embedding captures corresponding intent representation automatically. The recognition network $q_r(z|X)$ does not need additional training. We utilize the BERT intent classifier mentioned above, which achieves $87.75\%$ accuracy in intent classification. In addition, as the sample operation easily brings noise for the intent representation learning when sampling a wrong intent, we use argmax operation to avoid the noise, the response reconstruction loss is:

$$\mathcal{L}_1 = -\log p(X|C, z_k), \quad z_k = \arg\max_{z_k} q_r(z_k|X), \tag{2}$$

$k \in \{0, 1, 2, \cdots, 8\}$, each integer corresponds to a specific empathetic intent as in Figure 2.

### 4.3 Intent Predictor and Emotion Classifier

The intent predictor is based on the prior network $p_i(z|C)$, which predicts the distribution of response intent by the given context. During inference, we sample potential intents from this distribution in order to generate human-like empathetic responses. Specifically, the context is encoded with gated recurrent units (GRU) (Chung et al., 2014):

$$h_t = \mathbf{GRU}(h_{t-1}, E(c_t)), \tag{3}$$

where $h_t$ is the hidden state of GRU encoder, $E(c_t)$ denotes the word embedding of the $t$-th word in context, we use $h_m$ as context embedding, then the prior network is:

$$p_i(z|C) = \mathbf{Softmax}(\mathbf{FFN_z}(h_m)), \tag{4}$$

where $\mathbf{FFN}$ represents *Feed-Forward Network* with two layers. The prior intent distribution is supervised by recognition distribution with KL-divergence in Eq.1:

$$\begin{aligned} \mathcal{L}_2 &= \mathbf{KL}(q_r(z|X)||p_i(z|C)) \\ &= \sum_{k=1}^{K} q_r(z_k|X) \log \frac{q_r(z_k|X)}{p_i(z_k|C)}. \end{aligned} \tag{5}$$

Since the context emotion is proved to be beneficial to empathetic dialogue generation (Rashkin et al., 2019; Lin et al., 2019; Majumder et al., 2020), we also employ an emotion classifier to classify the emotion situation of context:

$$\begin{aligned} \mathcal{P} &= \mathbf{Softmax}(\mathbf{FFN_e}(h_m))], \\ p_{e_i} &= \mathcal{P}[i] \end{aligned} \tag{6}$$

Given the ground truth emotion label $\mathbf{e_t}$, the emotion classifier is trained with cross-entropy loss:

$$\mathcal{L}_3 = -\log p_{\mathbf{e_t}}. \tag{7}$$

### 4.4 Response Generator

As for the response generation $p(X|C, z)$, we consider implicit intent embedding for the high-level abstraction of an intent. In addition, we also introduce intent keywords for explicitly utilizing intent knowledge during the generation process.

**Implicit.** To generate response with an empathetic intent, the most intuitive approach is taking the intent embedding as additional input to decoder during the generation process. We also consider emotion embedding as traditional empathetic dialogue models:

$$s_t = \mathbf{GRU}(s_{t-1}, [E(x_{t-1}); v(z); v(e); c_{att}]), \quad (8)$$

where $s_t$ is the state of GRU decoder, $c_{att}$ denotes the context attention value which contains key information of context (Bahdanau et al., 2015). $v(z)$ is intent embedding and $v(e)$ is emotion embedding, both will not change during the generation process. However, this may sacrifice grammatical correctness (Zhou et al., 2018; Ghosh et al., 2017). Therefore we add a gate operation to capture intent and emotion dynamically:

$$\begin{aligned} \text{Input} &= \mathbf{FFN_i}([E(x_t); c_{att}; s_t]), \\ \text{Gate} &= \mathbf{Sigmoid}(\text{Input}), \\ \bar{v}(z) &= \text{Gate} \odot v(z), \end{aligned} \quad (9)$$

where $\odot$ represents element-wise product. Each time step, the intent representation is used appropriately according to current word, state, and context value, the gate operation is the same for emotion.

**Explicit.** The empathetic expression is quite distinct over vocabularies, e.g., ‘know’, ‘understand’, ‘agree’, are indicative of the empathetic intent *Agreeing*. Therefore, we employ the copy mechanism to explicitly utilize intent keywords for intent conditional generation. See Appendix for more details about intent keywords .

$$\begin{aligned} \alpha_t &= \mathbf{Sigmoid}(v_s^\top s_t), \\ p(x_t = w_g) &= \mathbf{Softmax}(W_g s_t), \\ p(x_t = w_i) &= \mathbf{Softmax}(W_i s_t), \\ p(x_t) &= (1 - \alpha_t) \cdot p(w_g) + \alpha_t \cdot p(w_i), \end{aligned} \quad (10)$$

where $\{s_t, v_s\} \in \mathcal{R}^{d\times 1}$, $\{W_g, W_i\} \in \mathcal{R}^{V\times d}$, $d$ is hidden size and $V$ denotes the vocabulary size. The copy rate $\alpha_t$ is used to balance the choice between intent keywords and generic words, it is trained with binary cross entropy loss:

$$\mathcal{L}_4 = \sum_{t=1}^{n} q_t \cdot \log \alpha_t + (1 - q_t) \cdot \log(1 - \alpha_t), \quad (11)$$

$n$ is the word number of response, $q_t \in \{0, 1\}$ is the truth whether the response word $x_t$ is intent keyword.

### 4.5 Loss Function

To summarize, the total loss is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 + \lambda_4 \mathcal{L}_4, \quad (12)$$

where $\lambda$ is the hyper-parameter controlling the proportion of its part. $\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3, \mathcal{L}_4$ denote the losses of response reconstruction, intent prediction, emotion classification and copy rate prediction respectively.

## 5 Experiments

### 5.1 Dataset

We evaluate our method and compare with others on **EmpatheticDialogues**[2] (Rashkin et al., 2019) which contains $25k$ open domain dialogues. Follow the same setting as the authors of this dataset, the proportion of train/validation/test data is $8:1:1$. Each dialogue consists of at least two utterances between a speaker and listener. There are 32 emotion situations in total, which are uniformly distributed.

### 5.2 Baselines

We compare our model with the three latest empathetic conversational models:

- **Multitask Transformer (Multi-TRS).** A transformer model trained by the response generation task and the context emotion classification task (Rashkin et al., 2019).
- **Mixture of Empathetic Listeners (MoEL).** An enhanced transformer model with 32 emotion-specific decoders to respond appropriately for each emotion (Lin et al., 2019).
- **MIMicking Emotions for Empathetic Response Generation (MIME).** The state-of-the-art empathetic dialogue model allows the generator to mimic the context emotion to a varying degree based on its positivity, negativity, and content. Furthermore, they introduce

[2] https://github.com/facebookresearch/EmpatheticDialogues

stochasticity into the emotion mixture and achieves one-to-many generation (Majumder et al., 2020).

## 5.3 Evaluation

### 5.3.1 Automatic Metrics

- **BLEU.** We choose BLEU (Papineni et al., 2002) for relevance evaluation which measures the $n$-gram overlaps with reference and compute BLEU scores for $n \leq 4$ using smoothing techniques (Chen and Cherry, 2014). Since the state-of-art model MIME and ours are both one-to-many generators, we calculate BLEU recall and BLEU precision (Zhao et al., 2017; Gu et al., 2019). For each test case, we sample $5$ responses from latent space and use greedy search for MIME and EmpHi, use beam search for MoEL and Multitask-Transformer.

- **Distinct.** Distinct (Li et al., 2016) is a widely used metric for diversity evaluation. Specifically, we compute the number of distinct unigrams (Distinct-1) and bigrams (Distinct-2), then scale them by the total number of unigrams and bigrams.

### 5.3.2 Human Ratings

First, we randomly sample 100 dialogues and their corresponding generations from the three baseline models and EmpHi. Then, we invite five volunteers with master degrees to do the human evaluation. The annotators mark each response from 1 to 5 for empathy, relevance, and fluency.

To clarify the marking criteria, we provide an explanation for each metric:

- **Empathy.** Whether the response shows that the listener understands and shares the speaker's feeling. Can the listener imagine what it would be like in the speaker's situation?

- **Relevance.** Whether the response is relevant to the context.

- **Fluency.** Whether the response is easy to read and grammatically correct.

### 5.3.3 Human A/B Test

Following (Lin et al., 2019; Majumder et al., 2020), we construct this evaluation task to directly compare our model with each baseline. We randomly sample $100$ dialogue responses from *EmpHi* vs {*Multitask-Trans*, *MoEL*, *MIME*}. Given randomly ordered responses from above models, four annotators select the better response, or $tie$ if they think the two responses have the same quality. The average score of four results is calculated, and shown in Table 6.

## 5.4 Implement Detail

For MIME[3] (Majumder et al., 2020) and MoEL[4] (Lin et al., 2019), we reproduce their results using their open-source codes and their default hyperparameters. According to the log-likelihood in the validation dataset for Multitask-Transformer, we use grid search for the best head number, layer number, and feed-forward neural network size. The best set is 2, 10, and 256, respectively. EmpHi uses a two-layer Bi-GRU as the encoder and a two-layer GRU as the decoder, $\lambda$ is set as $[1, 0.5, 0.5, 1]$ respectively. All the feed-forward neural networks in EmpHi have two layers, 300 hidden units and ReLU activations. For the sake of fairness, we use pretrained Glove vectors (Pennington et al., 2014) with 300 dimensions as the word embedding for all models, the batch size is 16, and the learning rate is set to $1e^{-4}$.

# 6 Results and Discussions

## 6.1 Results Analysis

In this section, we mainly testify:

- human-like empathetic intent boost EmpHi's performance in terms of empathy, relevance, and diversity.

- EmpHi successfully captures the empathetic intent distribution of humans.

### 6.1.1 Human Evaluation

As shown in Table 1, EmpHi outperforms all baselines in terms of empathy, relevance, and fluency. The most distinct improvement is 15.1% on relevance because our model does not only depends on the speakers' emotion, but also makes use of the empathetic intents, which are contextually relevant. It is worth noting that empathy is the primary metric in empathetic dialogue generation. EmpHi outperforms the previous SOTA on empathy by 9.43%, which directly indicates that human-like empathetic intents are beneficial to the empathy

[3] https://github.com/declare-lab/MIME
[4] https://github.com/HLTCHKUST/MoEL

| Methods | #Params. | Empathy | Relevance | Fluency | BLEU | | | Distinct | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | P | R | F1 | D-1 | D-2 |
| Multitask-Trans | 20M | 2.68 | 2.58 | 3.60 | 0.3072 | 0.4137 | 0.3526 | 0.4123 | 1.1390 |
| MoEL | 21M | 3.18 | 3.18 | 3.95 | 0.3032 | 0.3614 | 0.3298 | 0.8473 | 4.4698 |
| MIME | 18M | 2.89 | 2.90 | 3.77 | 0.3202 | 0.3278 | 0.3240 | 0.3952 | 1.3299 |
| EmpHi | **11M** | **3.48** | **3.66** | **4.34** | **0.3207** | **0.4723** | **0.3820** | **1.1188** | **5.3332** |
| Human | - | 4.04 | 4.40 | 4.56 | - | - | - | 7.0356 | 43.2174 |

Table 1: Automatic evaluation between EmpHi and other models. All results are the mean of 5 runs for fair comparison. D-1.&2. are magnified 100 times for each model.

| Methods | Win | Loss | Tie |
|---|---|---|---|
| EmpHi vs Multitask-trans | **56.5%** | 21.5% | 22.0% |
| EmpHi vs MoEL | **45.0%** | 28.5% | 26.5% |
| EmpHi vs MIME | **53.0%** | 27.0% | 20.0% |

Table 2: Results of Human A/B test.

| Methods | BLEU | | | ACC |
|---|---|---|---|---|
| | P | R | F1 | |
| EmpHi | 0.3207 | **0.4723** | **0.3820** | **26.8%** |
| EmpHi w/o intent | 0.3105 | 0.4049 | 0.3515 | 21.9% |
| EmpHi w/o gate | 0.3138 | 0.4634 | 0.3742 | 25.3% |
| EmpHi w/o copy | **0.3215** | 0.4704 | 0.3820 | 25.9% |

Table 3: Results of ablation study.

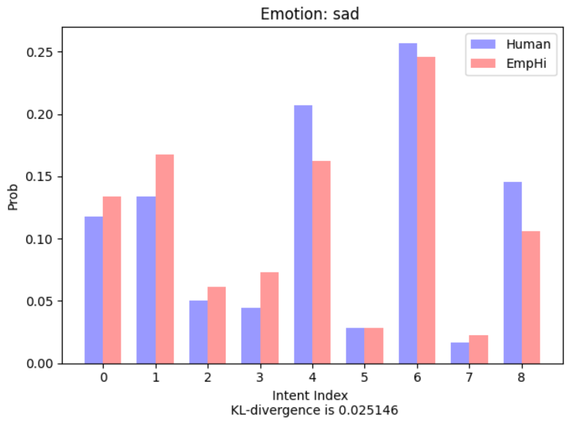


Figure 6: Empathetic intent distribution of human and EmpHi (sad emotion), the intent index represents the same intent as in Figure 2.

ability of the dialogue model. Last but not least, a decent fluency score proves that our generated response could be understood by humans easily, where our model shows an improvement of 9.87%. In addition, the human A/B test results in Table 2 also confirm that the responses from our model are preferable to baselines. Overall, EmpHi successfully generates empathetic, relevant, and fluent responses.

#### 6.1.2 Automatic Evaluation

As seen in Table 1, the automatic evaluation is consistent with human evaluation. The BLEU recall and F1 score are improved by 14.2% and 8.34%, respectively. However, we only have a slight improvement on BLEU precision, which is similar to (Zhao et al., 2017; Gu et al., 2019) because the precision is penalized when the model generates diverse responses. Also, the distinct value of unigrams and bigrams are higher, about 32.04% and 19.32% than the previous SOTA, respectively. As shown in Figure 4 and Figure 6, the empathy intents of EmpHi's responses are more diverse than existing models, so the distinct scores improve significantly. It should be noticed that our method enhances the relevance and diversity simultaneously, which proves the effectiveness of human-like intent in empathetic response generation.

#### 6.1.3 Empathetic Intent Distribution

We apply the same approach in Section 3 and quantify the empathetic intent distribution of EmpHi's responses to prove that EmpHi accurately captures humans' empathetic intent distribution. Comparing Figure 4 and Figure 6, the difference between them illustrates that our model successfully reduces the bias of empathetic expression. The KL-divergence of intent distributions between models and humans are **0.025** for EmpHi, 1.949 for MIME, 1.545 for MoEL, and 4.570 for Multitask-Transformer (See in Appendix).

### 6.2 Ablation Study

We evaluate each part of EmpHi according to BLEU and ACC, where ACC indicates the accuracy of whether the empathetic intent of generated response is the same as ground truth. This value is low because one context could have multiple empathetic responses with various empathetic intents. As seen in Table 3, there is a dramatic drop in the performance of EmpHi without any intent information (both implicit embedding and explicit keywords). Therefore, this proves the effectiveness

| | |
|---|---|
| **Emotion** | Nostalgic |
| **Context** | When my wife and i started dating, our second date ended up lasting like 12 hours. |
| **Reference** | That is a pretty long date ! Where did you guys go? |
| **Multitask-Trans** | That is so sad, I'm sorry to hear that. |
| **MoEL** | That is great! I am sure she will be fine. |
| **MIME** | That is a good thing to do. I am sure you will do great. |
| **EmpHi** | That 's awesome ! How long have you been together ? |
| **Emotion** | Angry |
| **Turn-1** | I just moved to this neighborhood and some dumb criminals shot one of my neighbors and ran into the woods! |
| **Turn-2** | That's not good. Do you own a gun? |
| **Turn-3** | I do! I want to be able to protect my son. |
| **Reference** | That is always number one goal. |
| **Multitask-Trans** | What did you do ? |
| **MoEL** | That is a good idea . |
| **MIME** | I am sorry to hear that. I hope you get it ! |
| **EmpHi** | Maybe you should go to the police. |

Table 4: Comparison among EmpHi and other empathetic dialogue models.

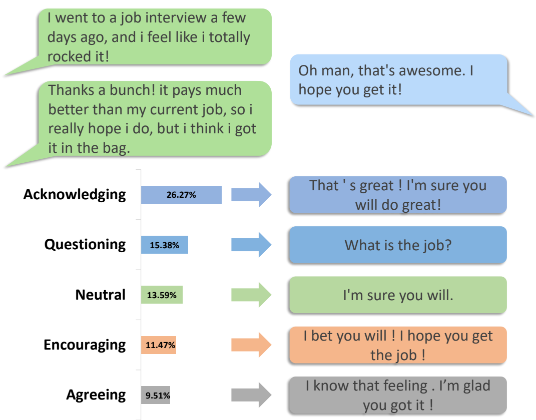


Figure 7: Case study of EmpHi.

of empathetic intents and the intent representation learning approach. As for implicit gate control, we can tell that it improves both response quality and intent accuracy because it helps dynamically capture intent information during generation. This result is the same as (Zhou et al., 2018). The copy mechanism provides EmpHi the ability to explicitly use intent keywords and thus contributes to the intent accuracy.

### 6.3 Case Study

**Intent-level diverse generation.** Through sampling intents in the discrete latent space, EmpHi generates different responses with empathetic intents. As in Figure 7, the speaker shows an exciting emotion for the opportunity of getting a better job. EmpHi generates empathetic yet contextually relevant responses as humans. Besides, EmpHi predicts the potential intent distribution and shows successful conditional generation based on the corresponding intents, which improves the interpretability and controllability of empathetic response generation. We also have an error analysis for a more comprehensive understanding of EmpHi in the Appendix.

**Compare with existing models.** For the first instance in Table 4, even though baseline models show naive empathy in their response, it is hard for the speaker to feel empathy because the response is not relevant to the topic. In contrast, EmpHi shows its understanding of the speaker's feelings and asks a relevant question to explore the speaker's experience. The second case tells the same story. Again, all baselines express contextually irrelevant empathy, whereas EmpHi truly understands the scene and further reply: "Maybe you should go to the police" with the *Suggesting* intent.

## 7 Conclusion

Overall, we reveal the severe bias of empathetic expression between existing dialogue models and humans. To address this issue, this paper proposes EmpHi to generate empathetic responses with human-like empathetic intents. As a result, both automatic and human evaluation prove that EmpHi has a huge improvement on empathetic conversation. According to the anlaysis and case studies, EmpHi successfully learns to be empathetic consistently with humans and shows high interpretability during the generation process.

We will add more empathetic intents, e.g., delighting, cheering, persuading, etc, and try large pretrained models in our future work.

## 8 Ethical Statement

Since this paper involves subjects related to human conversation, we have ensured that all the experiments will cause no harm to humans. The dataset EmpatheticDialogues is collected by (Rashkin et al., 2019), all the participants join the data collection voluntarily. Also, the dataset provider filters all personal information and obscene languages. Therefore, we believe that the dataset EmpatheticDialogues used in our experiments are harmless to users, and the model trained on this dataset is not dangerous to humans.

## References


Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21. ACL.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4758–4765. Association for Computational Linguistics.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 362–367. The Association for Computer Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 634–642. Association for Computational Linguistics.

Xiaodong Gu, Kyunghyun Cho, Jung-Woo Ha, and Sunghun Kim. 2019. Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 121–132. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8968–8979. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW*.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5263–5276. Association for Computational Linguistics.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4886–4899. International Committee on Computational Linguistics.

Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. 2016. Attribute2image: Conditional image generation from visual attributes. 9908:776–791.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1098–1107. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.