Towards Persona-Based Empathetic Conversational Models

Peixiang Zhong[1,2,3], **Chen Zhang**[4], **Hao Wang**[4], **Yong Liu**[1,3], **Chunyan Miao**[1,2,3*]
[1]Alibaba-NTU Singapore Joint Research Institute
[2]School of Computer Science and Engineering
[3]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly
Nanyang Technological University, Singapore
[4]Alibaba Group, China
peixiang001@e.ntu.edu.sg, zhangchen010295@163.com, cashenry@126.com,
stephenliu@ntu.edu.sg, ascymiao@ntu.edu.sg

Abstract

Empathetic conversational models have been shown to improve user satisfaction and task outcomes in numerous domains. In Psychology, persona has been shown to be highly correlated to personality, which in turn influences empathy. In addition, our empirical analysis also suggests that persona plays an important role in empathetic conversations. To this end, we propose a new task towards persona-based empathetic conversations and present the first empirical study on the impact of persona on empathetic responding. Specifically, we first present a novel large-scale multi-domain dataset for persona-based empathetic conversations. We then propose CoBERT, an efficient BERT-based response selection model that obtains the state-of-the-art performance on our dataset. Finally, we conduct extensive experiments to investigate the impact of persona on empathetic responding. Notably, our results show that persona improves empathetic responding more when CoBERT is trained on empathetic conversations than non-empathetic ones, establishing an empirical link between persona and empathy in human conversations.

1 Introduction

Empathy, specifically affective empathy, refers to the capacity to respond with an appropriate emotion to another's mental states (Rogers et al., 2007). In NLP, empathetic conversational models have been shown to improve user satisfaction and task outcomes in numerous domains (Klein, 1998; Liu and Picard, 2005; Wright and McCarthy, 2008; Fitzpatrick et al., 2017; Zhou et al., 2018a). For example, empathetic agents received more positive user ratings, including greater likeability and trustworthiness than controls (Brave et al., 2005).

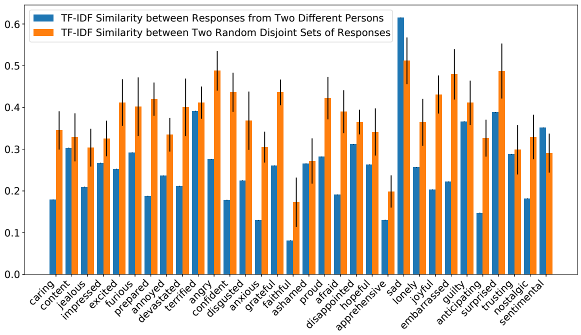

Figure 1: TF-IDF similarity between two sets of empathetic responses (Rashkin et al., 2019) for each emotion (best viewed in color). For most emotions (28 out of 32), the similarity between responses from two different speakers (blue) is substantially smaller than the similarity between two random disjoint sets of responses (orange, averaged over five runs).

In recent years, neural network based conversational models (Vinyals and Le, 2015; Lowe et al., 2015) are becoming dominant. Zhou et al. (2018a) designed XiaoIce, a popular AI companion with an emotional connection to satisfy the human need for communication, affection, and social belonging. Recently, Rashkin et al. (2019) presented a new dataset and benchmark towards empathetic conversations and found that both Transformer-based generative models (Vaswani et al., 2017) and BERT-based retrieval models (Devlin et al., 2019) relying on this dataset exhibit stronger empathy.

However, most existing studies, e.g., (Rashkin et al., 2019), do not consider persona when producing empathetic responses[1]. In Psychology, persona refers to the social face an individual presents to the world (Jung, 2016). Persona has been shown to be highly correlated with personality (Leary and Allen, 2011), which in turn influences empathy (Richendoller and Weaver III, 1994; Costa et al., 2014). In addition, our empirical analysis of empathetic con-

[*] Corresponding author

[1] One exception is XiaoIce (Zhou et al., 2018a), however, her persona is not configurable and thus difficult to satisfy various human needs.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6556–6566,
November 16–20, 2020. ©2020 Association for Computational Linguistics

versations in (Rashkin et al., 2019) also shows that for most emotions, the empathetic responses from two different persons[2] have more differences than that between two disjoint sets of random responses, as shown in Figure 1. Both the theories in Psychology and the evidence from our empirical analysis suggest that persona plays an important role in empathetic conversations, which, to the best of our knowledge, has not been investigated before[3].

To this end, we propose a new task towards persona-based empathetic conversations and present the first empirical study on the impact of persona on empathetic responding. Our study would be beneficial to researchers in Dialogue Systems and Psycholinguistics. However, one major challenge of this study is the lack of relevant datasets, i.e., existing datasets only focus on either persona or empathy but not both (see Table 4 for details). In this paper, we present a novel large-scale multi-turn **P**ersona-based **E**mpathetic **C**onversation (**PEC**) dataset in two domains with contrasting sentiments, obtained from the social media Reddit, to facilitate our study.

We then propose CoBERT, an efficient BERT-based response selection model using multi-hop co-attention to learn higher-level interactive matching. CoBERT outperforms several competitive baselines on PEC, including Poly-encoder (Humeau et al., 2020), the state-of-the-art BERT-based response selection model, by large margins. We conduct additional comparisons with several BERT-adapted models and extensive ablation studies to evaluate CoBERT more comprehensively.

Finally, based on PEC and CoBERT, we investigate the impact of persona on empathetic responding. In addition, we analyze how limited persona data improves model performance, and how our model generalizes to new personas.

In summary, our contributions are as follows:

- We propose a new task and a novel large-scale multi-domain dataset, PEC, towards persona-based empathetic conversations. Our data and code are available here[4].

- We propose CoBERT, a BERT-based response selection model that obtains the state-of-the-art performance on PEC. Extensive experimental evaluations show that CoBERT is both effective and efficient.

- We present the first empirical study on the impact of persona on empathetic responding. The results show that persona improves empathetic responding *more* when CoBERT is trained on empathetic conversations than non-empathetic ones, establishing an empirical link between persona and empathy in human conversations.

2 Related Work

Empathetic Conversational Models Despite the growing number of studies in neural conversational models, less attention has been paid to make conversations empathetic until recently (Siddique et al., 2017; Morris et al., 2018; Shi and Yu, 2018; Lin et al., 2019b; Shin et al., 2019; Rashkin et al., 2019; Li et al., 2019; Lin et al., 2019a; Zandie and Mahoor, 2020), possibly due to the lack of empathetic conversation datasets. Rashkin et al. (2019) proposed EMPATHETICDIALOGUES (**ED**), the first empathetic conversation dataset comprising 25K conversations in 32 emotions. Conversational models trained on the role of the listener in the dataset exhibited stronger empathy than models trained on non-empathetic datasets. We compare ED and PEC in the last paragraph of Section 3.

Persona-Based Conversational Models In recent years, personalized conversational models are emerging (Li et al., 2016; Zhang et al., 2018a; Wolf et al., 2019; Chan et al., 2019; Madotto et al., 2019; Zheng et al., 2019). Li et al. (2016) proposed persona embeddings in a response generation model and achieved improved generation quality and persona consistency. Zhang et al. (2018a) proposed PERSONA-CHAT (**PC**), a crowd-sourced conversation dataset with persona information, to improve model engagingness and consistency. Mazare et al. (2018) further presented a much larger persona-based conversation dataset collected from Reddit (**PCR**) and showed that persona consistently improves model performance even when a large number of conversations is available for training. We compare PC, PCR, and PEC in the last paragraph of Section 3. Recently, Gu et al. (2019) proposed DIM, a personalized response selection model with interactive matching and hierarchical aggregation, and achieved state-of-the-art performance on PC.

Retrieval-based Conversational Models Recent neural retrieval-based conversational models gener-

[2] Each response in (Rashkin et al., 2019) has a speaker id but no persona.

[3] A very recent work (Roller et al., 2020) incorporates persona and empathy by fine-tuning on corresponding datasets, however, it does not investigate the impact of persona on empathetic responding.

[4] https://github.com/zhongpeixiang/PEC

	happy			offmychest		
	train	valid	test	train	valid	test
#Conv.	157K	20K	23K	124K	16K	15K
#Utter.	367K	46K	54K	293K	38K	35K
#Speaker	93K	17K	19K	89K	16K	16K
#Avg.PS	66.0	70.8	70.0	59.6	66.8	67.1
#Std.PS	38.1	36.7	36.9	40.2	39.0	38.8
#Avg.U	21.5	21.9	21.3	30.4	31.5	30.0
#Avg.P	10.9	10.8	10.8	10.9	10.9	10.9

Table 1: Statistics of PEC. #Avg.PS and #Std.PS denote average and standard deviation of the number of persona sentences per speaker, respectively. #Avg.U denotes the average utterance length. #Avg.P denotes the average persona sentence length.

	happy	offmychest	control group
Sentiment	0.85	-0.39	0.03
Empathy	0.73	0.61	0.25

Table 2: Sentiment and empathy of PEC and the control group based on human ratings. Sentiment ranges from -1 (negative) to 1 (positive). Empathy ranges from 0 (non-empathetic) to 1 (empathetic). Ratings are aggregated by majority voting (averaging shows similar results). The inter-annotator agreement, measured by Fleiss' kappa (Fleiss, 1971), for sentiment and empathy are 0.725 and 0.617, respectively. Both agreement statistics indicate "substantial agreement".

ally have three modules: encoding, matching and aggregation (Lowe et al., 2015; Zhou et al., 2016; Wu et al., 2017; Zhou et al., 2018b; Zhang et al., 2018b; Chen and Wang, 2019; Feng et al., 2019; Yuan et al., 2019). The encoding module encodes text into vector representations using encoders such as LSTM, Transformer, or BERT. The matching module measures context-response associations using various attention mechanisms at different granularities. The aggregation module summarizes the matching information along the sequence dimension to obtain the final representation. A recent work Humeau et al. (2020) proposed Poly-encoder, an efficient BERT-based response selection model that obtained the state-of-the-art performance on multiple conversation datasets.

3 The PEC Dataset

In this section, we introduce the collection procedure and statistics of our proposed persona-based empathetic conversation (PEC) dataset.

Data Source We collect empathetic conversations from two subreddits *happy*[5] and *offmychest*[6] on Reddit, a discussion forum where users can discuss any topics on their corresponding subforums/subreddits. The *happy* subreddit is where users share and support warm and happy stories and thoughts. The *offmychest* subreddit is where users share and support deeply emotional things that users cannot tell people they know. We choose these two subreddits as our data source because their posts have contrasting sentiments and their comments are significantly more empathetic than casual conversations, i.e., the control group, as shown in Table 2.

Conversation Collection Discussions on Reddit are organized in threads where each thread has one post and many direct and indirect comments. Each thread forms a tree where the post is the root node and all comment nodes reply to their parent comment nodes or directly to the root node. Therefore, given a thread with n nodes, we can extract $n-1$ conversations where each conversation starts from the root node and ends at the $n-1$ non-root nodes. We randomly split conversations by threads according to the ratio of 8:1:1 for training, validation, and test sets, respectively.

Persona Collection Following (Mazare et al., 2018), for each user in the conversations, we collect persona sentences from all posts and comments the user wrote on Reddit. The posts and comments are split into sentences, and each sentence must satisfy the following rules to be selected as a persona sentence: 1) between 4 and 20 words; 2) the first word is "i"; 3) at least one verb; 4) at least one noun or adjective; and 5) at least one content word. Our rules are stricter than that from (Mazare et al., 2018), allowing us to extract less noisy persona sentences. For each user, we extract up to 100 persona sentences.

Note that we choose our approach to persona collection because 1) the well-established work (Mazare et al., 2018) successfully trained personalized agents using this approach; 2) this approach is significantly more scalable and cost-effective than crowd-sourcing; and 3) we are concerned that using crowd-sourcing, i.e., assigning artificial personas to crowd-workers and asking them to chat empathetically based on the assigned personas, would introduce worker-related noises such that models may merely learn superficial empathetic responding patterns that crowd-workers deem suitable given the assigned personas.

Data Processing We keep a maximum of 6 most recent turns for each conversation. We filter con-

[5] https://www.reddit.com/r/happy/

[6] https://www.reddit.com/r/offmychest/

	happy	offmychest
Conversation	Celebrating 43 years of marriage with the love of my life.	Worried. Am I becoming depressed again? Please don't leave me. Is everything okay? You don't seem yourself.
	She looks very young for someone who has been married 43 years. That must surely put her in the 63-73yr age range?!	I'm living these exact words.
	I just turned 61, thanks!	I hope everything works out for you. I'm trying not to fall apart.
	I hope I look that young when I'm 61! You guys are too cute, congratulations :)	Me too. If you ever want someone to talk to my messages are open to you.
Persona	I took an 800 mg Ibuprofen and it hasn't done anything to ease the pain.	I think I remember the last time I ever played barbies with my litter sister.
	I like actively healthy.	I have become so attached to my plants and I really don't want it to die.
	I want a fruit punch!	I'm just obsessed with animals.

Table 3: Two example conversations with personas from PEC. The persona sentences correspond to the last speakers in the conversations.

Dataset	Source	Persona	Empathy	Size	Public
ED	CS	✗	✓	78K	✓
PC	CS	✓	✗	151K	✓
PCR	Reddit	✓	✗	700M	✗
PEC (ours)	Reddit	✓	✓	355K	✓

Table 4: Comparisons between PEC and related datasets. ED denotes EMPATHETICDIALOGUES (Rashkin et al., 2019). PC denotes PERSONA-CHAT (Zhang et al., 2018a). PCR denotes the persona-based conversations from Reddit (Mazare et al., 2018). CS denotes crowd-sourced. The size denotes the number of expanded conversations.

versations to ensure that 1) each post is between 2 and 90 words; 2) each comment is between 2 and 30 words[7]; 3) all speakers have at least one persona sentence; and 4) the last speaker is different from the first speaker in each conversation. The last requirement is to maximally ensure that the last utterance is the empathetic response instead of a reply of the poster. In addition, persona sentences appearing in the conversation responses are removed to avoid data leakage. Finally, we lowercase all data and remove special symbols, URLs, and image captions from each sentence. The statistics of PEC are presented in Table 1. Two examples of PEC are shown in Table 3.

Note that it may not be easy to see explicit links in Table 3, but that's exactly what we are studying for, i.e., to uncover the implicit (and possibly unexpected) links between persona and empathy using real user data. For example, the utterance "I hope I look that young" may implicitly link to the persona "I like actively healthy" in Table 3.

Data Annotations We manually annotate 100 randomly sampled conversations from each domain to estimate their sentiment and empathy. To avoid annotation bias, we add a control group comprising 100 randomly sampled casual conversations from the *CasualConversation*[8] subreddit, where users can casually chat about any topics. Finally, we mix and shuffle these 300 conversations and present them to three annotators. The annotation results are presented in Table 2. The posts in the happy and offmychest domains are mostly positive and negative, respectively. Both domains are significantly more empathetic than the control group ($p < 0.001$, one-tailed t-test).

Conversation Analysis We conduct conversation analysis for PEC, similar to our analysis for ED (Rashkin et al., 2019) in Figure 1. Specifically, the TF-IDF similarities between responses from two different persons are 0.25 and 0.17 for happy and offmychest, respectively, whereas the TF-IDF similarities between two disjoint sets of random responses are 0.38 (±0.05) and 0.31 (±0.05) for happy and offmychest over 5 runs, respectively. The results show that empathetic responses between different persons are more different than that between random empathetic responses in PEC, suggesting that different speakers in PEC have different "styles" for empathetic responding.

Comparisons with Related Datasets Table 4 presents the comparisons between PEC and related datasets. PEC has the unique advantage of being both persona-based and empathetic. In addition, PEC is collected from social media, resulting in a much more diverse set of speakers and language patterns than ED (Rashkin et al., 2019) and PC (Zhang et al., 2018a), which are collected from only hundreds of crowd-sourced workers. Finally,

[7] Posts are usually longer than comments. 87% posts and 82% comments on *happy* are less than 90 and 30 words, respectively. 24% posts and 59% comments on *offmychest* are less than 90 and 30 words, respectively.

[8] https://www.reddit.com/r/CasualConversation/

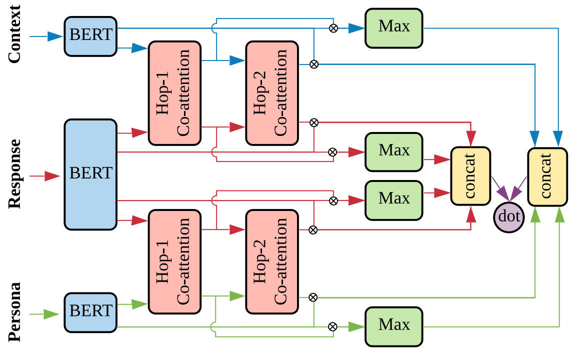

Figure 2: Our CoBERT architecture.

PEC is over 2x larger than the other two public datasets, allowing the exploration of larger neural models in future research.

4 Our CoBERT Model

In this section, we briefly introduce the task of response selection and present our proposed CoBERT model, as shown in Figure 2.

4.1 Task Definition

We denote a training conversation dataset $\mathcal{D}$ as a list of N conversations in the format of (X, P, y), where $X = \{X_1, X_2, ..., X_{n_X}\}$ denotes the n_X context utterances, $P = \{P_1, P_2, ..., P_{n_P}\}$ denotes the n_P persona sentences of the respondent, and y denotes the response to X. The task of response selection can be formulated as learning a function $f(X, P, y)$ that assigns the highest score to the true candidate y and lower scores to negative candidates given X and P. During inference, the trained model selects the response candidate with the highest score from a list of candidates.

4.2 BERT Representation

We use BERT (Devlin et al., 2019) as our sentence encoders. Similar to the Bi-encoder (Humeau et al., 2020), we concatenate context utterances as a single context sentence before passing it into BERT. Since there is no ordering among persona sentences, we concatenate randomly ordered persona sentences[9]. After passing the context, persona and response to BERT encoders, we obtain their vector representations $\mathbf{X} \in \mathbb{R}^{m \times d}$, $\mathbf{P} \in \mathbb{R}^{q \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d}$ from the last layer, respectively, where d denotes the embedding size of BERT, and m, q and n denote the sequence lengths of context, persona and response, respectively. Note that different segment ids are used to differentiate speaker and respondent utterances in the context.

4.3 Hop-1 Co-attention

Given $\mathbf{X}$ and $\mathbf{Y}$, we learn the first-order matching information using co-attention (Lu et al., 2016). Specifically, we first compute the word-word affinity matrix $\mathbf{A_{XY}} \in \mathbb{R}^{m \times n}$:

$$\mathbf{A_{XY}} = \mathbf{X}\mathbf{Y}^T. \tag{1}$$

Then the context-to-response attention $\mathbf{A_{X2Y}} \in \mathbb{R}^{m \times n}$ and the response-to-context attention $\mathbf{A_{Y2X}} \in \mathbb{R}^{n \times m}$ can be computed as follows:

$$\mathbf{A_{X2Y}} = \mathit{softmax}(\mathbf{A_{XY}}), \tag{2}$$

$$\mathbf{A_{Y2X}} = \mathit{softmax}(\mathbf{A_{XY}^T}), \tag{3}$$

where *softmax* denotes the softmax function along the second dimension. Finally, we obtain the attended context representation $\mathbf{X}' = \mathbf{A_{X2Y}}\mathbf{Y} \in \mathbb{R}^{m \times d}$ and response representation $\mathbf{Y_X'} = \mathbf{A_{Y2X}}\mathbf{X} \in \mathbb{R}^{n \times d}$.

To aggregate the first-order matching information and extract discriminative features, we apply max-pooling to $\mathbf{X}'$ and $\mathbf{Y_X'}$ along the sequence dimension and obtain $\mathbf{X}'_{max} \in \mathbb{R}^d$ and $\mathbf{Y}'_{\mathbf{X},max} \in \mathbb{R}^d$.

4.4 Hop-2 Co-attention

We propose a hop-2 co-attention to learn second-order interactive matching. Different from the attention-over-attention for reading comprehension (Cui et al., 2017), our method learns bidirectional matching for response selection. Specifically, we apply attention over the attention matrices:

$$\mathbf{A_X}' = \mathit{mean}(\mathbf{A_{X2Y}})\mathbf{A_{Y2X}}, \tag{4}$$

$$\mathbf{A_Y}' = \mathit{mean}(\mathbf{A_{Y2X}})\mathbf{A_{X2Y}}, \tag{5}$$

where $\mathbf{A_X}' \in \mathbb{R}^{1 \times m}$ and $\mathbf{A_Y}' \in \mathbb{R}^{1 \times n}$ denote the second-order attention over $\mathbf{X}$ and $\mathbf{Y}$, respectively, and *mean* denotes mean pooling along the first dimension. Then we obtain the attended context representation $\mathbf{X}'' = \mathbf{A_X}'\mathbf{X} \in \mathbb{R}^d$ and response representation $\mathbf{Y_X''} = \mathbf{A_Y}'\mathbf{Y} \in \mathbb{R}^d$.

We apply the same procedure to match $\mathbf{P}$ and $\mathbf{Y}$, and obtain the first-order matching information $\mathbf{P}'_{max} \in \mathbb{R}^d$ and $\mathbf{Y}'_{\mathbf{P},max} \in \mathbb{R}^d$, and the second-order matching information $\mathbf{P}'' \in \mathbb{R}^d$ and $\mathbf{Y_P''} \in \mathbb{R}^d$.

[9] Reusing the same positional information for all persona sentences (Wolf et al., 2019) to model position invariance produces worse performance in our preliminary experiments.

Intuitively, our hop-1 co-attention learns attended representations for $\mathbf{X}$ and $\mathbf{Y}$, and our hop-2 co-attention learns "truly" attended representations for $\mathbf{X}$ and $\mathbf{Y}$ where the weights are computed from attentions over attentions.

4.5 Loss

We obtain the final persona-aware context representation $\mathbf{X}_f = [\mathbf{X}'_{max}; \mathbf{X}''; \mathbf{P}'_{max}; \mathbf{P}''] \in \mathbb{R}^{4d}$ and the final response representation $\mathbf{Y}_f = [\mathbf{Y}'_{\mathbf{X},max}; \mathbf{Y}''_{\mathbf{X}}; \mathbf{Y}'_{\mathbf{P},max}; \mathbf{Y}''_{\mathbf{P}}] \in \mathbb{R}^{4d}$, where $[;]$ denotes concatenation. Then we use dot product to compute the final matching score:

$$f(X, P, y) = dot(\mathbf{X}_f, \mathbf{Y}_f). \quad (6)$$

We optimize our model by minimizing the cross-entropy loss for selecting the true candidate from a list of candidates. Formally, the loss Φ is computed as follows:

$$\Phi = \sum_{(X,P,y)\sim\mathcal{D}} -\frac{e^{f(X,P,y)}}{\sum_{\hat{y}\sim\mathcal{N}(X)\cup\{y\}} e^{f(X,P,\hat{y})}}, \quad (7)$$

where $\mathcal{N}(X)$ denotes a set of randomly sampled negative candidates for the context X.

5 Experiments

In this section we present the datasets, baselines, experimental settings, model comparisons and ablation studies.

5.1 Datasets and Baselines

We evaluate models on PEC and its two subdomains, i.e., happy and offmychest. The training, validation and test splits of PEC are combined from the corresponding splits from happy and offmychest. The dataset statistics are shown in Table 1.

We compare CoBERT with several competitive baselines. Note that the BoW, HLSTM (Lowe et al., 2015) and Bi-encoder (Humeau et al., 2020) baselines share the same Tri-encoder architecture, where the final matching score is the dot product between the average of context and persona representations and the response representation.

BoW: The context, persona and response encoders compute the averaged word embedding.

HLSTM (Lowe et al., 2015): The context encoder has an utterance-level BiLSTM and a context-level BiLSTM. All encoders share the same utterance-level BiLSTM.

DIM (Gu et al., 2019): A state-of-the-art non-pretraiend model for persona-based response selection. DIM adopts finer-grained matching and hierarchical aggregation to learn rich matching representation.

Bi-encoder (Humeau et al., 2020): A state-of-the-art BERT-based model for empathetic response selection (Rashkin et al., 2019).

Poly-encoder (Humeau et al., 2020): A state-of-the-art BERT-based model for response selection. Poly-encoder learns latent attention codes for finer-grained matching. Note that we do not consider Cross-encoder (Humeau et al., 2020) as an appropriate baseline because it performs two orders of magnitude slower than Poly-encoder in inference, rendering it intractable for real-time applications.

5.2 Experimental Settings

Model Settings We use fastText (Paszke et al., 2019) embeddings of size 300 to initialize BoW and HLSTM. We follow the released code[10] to implement DIM. For all BERT-based models, we use the base version of BERT and share parameters across all three encoders[11]. We use 128 context codes for Poly-encoder[12]. We optimize all BERT-based models using Adam (Kingma and Ba, 2014) with batch size of 64 and learning rate of 0.00002. The positive to negative candidates ratio during training is set to 1:15. We use a maximum of $n_X = 6$ contextual utterances and a maximum of $n_P = 10$ persona sentences for each conversation. We conduct all experiments on NVIDIA V100 32GB GPUs in mixed precision.

Evaluation Metrics Following (Zhou et al., 2018b; Gu et al., 2019; Humeau et al., 2020), we evaluate models using Recall@k where each test example has C possible candidates to select from, abbreviated to R@k, as well as mean reciprocal rank (MRR). In our experiments, we set $C = 100$ and $k = 1, 10, 50$. The candidate set for each test example includes the true response and other $C - 1$ randomly sampled responses from the test set.

5.3 Comparison with Baselines

We report the test results of response selection in Table 5. Among the non-pretrained models, DIM

[10] https://github.com/JasonForJoy/DIM

[11] A shared BERT encoder obtained better performance than separate encoders in our preliminary experiments.

[12] More context codes result in memory error in our experiments. According to (Humeau et al., 2020), more context codes only lead to marginally better results.

	happy				**offmychest**				**PEC (happy + offmychest)**			
Models	R@1	R@10	R@50	MRR	R@1	R@10	R@50	MRR	R@1	R@10	R@50	MRR
BoW	10.2	45.6	85.2	21.8	13.9	51.6	87.1	26.2	15.4	52.9	86.7	27.4
HLSTM	15.7	53.6	91.6	28.1	17.6	55.7	91.8	30.2	22.2	63.0	94.8	35.2
DIM	31.3	67.0	95.5	43.0	40.6	72.6	96.4	51.2	39.3	74.6	97.3	50.5
Bi-encoder	32.4	71.3	96.5	45.1	42.4	78.4	97.6	54.5	42.3	79.2	98.1	54.4
Poly-encoder	33.7	72.1	96.7	46.4	43.4	79.3	97.7	55.3	43.0	79.8	98.2	55.2
CoBERT (ours)	**36.2**	**73.0**	**96.9**	**48.4**	**47.0**	**79.7**	**97.8**	**58.0**	**45.1**	**80.5**	**98.3**	**56.7**

Table 5: Test performance (in %) of CoBERT and all baselines. Values in bold denote best results.

Test Train	happy	offmychest	PEC
happy	36.2	41.2	40.5
offmychest	28.8	47.0	38.4
PEC	37.0	47.5	45.1

Table 6: Transfer test of CoBERT in R@1 (in %).

Model	R@1	MRR	InfTime	RAM
Baselines				
DIM	40.3	51.6	10.36x	**0.79x**
Bi-encoder	42.6	55.2	1.00x	1.00x
Poly-encoder	43.3	55.7	1.33x	1.84x
BERT-adapted Models				
BERT+MemNet	42.3	53.8	**0.87x**	0.89x
BERT+DAM	45.0	56.9	14.26x	1.57x
BERT+DIM	46.1	57.7	18.36x	1.78x
Ablations				
CoBERT (ours)	**46.2**	**57.9**	2.00x	1.14x
- hop-1	44.0	56.2	1.65x	1.11x
- hop-2	45.5	57.1	1.76x	1.11x
+ hop-3	46.0	57.6	2.70x	1.13x
- max + mean	44.1	56.3	2.12x	1.13x
+ mean	46.1	57.8	2.71x	1.15x

Table 7: Validation performance (in %), inference time (InfTime) and memory usage (RAM) for baselines, BERT-adapted models and ablation studies on PEC. InfTime and RAM are relative to the Bi-encoder.

outperforms BoW and HLSTM by large margins on all datasets, demonstrating the importance of finer-grained matching and hierarchical aggregation for response selection. The simple Bi-encoder performs noticeably better than DIM, suggesting that sentence representation is another critical factor in response selection and that BERT can provide much richer representation than the BiLSTM used in DIM. Poly-encoder performs best among all baselines because it leverages the strengths of both BERT and attention-based finer-grained matching.

Our CoBERT consistently outperforms all baselines on all datasets with large margins, including the state-of-the-art Poly-encoder. The performance gain is primarily attributed to our multi-hop co-attention, which learns higher-order bidirectional word-word matching between context and response, whereas Poly-encoder only learns the first-order unidirectional attention from response to context using latent attention codes. Efficiency-wise, CoBERT has slightly longer inference time (1.50x) but requires much less memory usage (0.62x) than Poly-encoder, as shown in Table 7.

We further investigate the transfer performance of CoBERT in Table 6. In general, in-domain test results are better than out-of-domain test results. The transfer performance from happy to offmychest (41.2%) and vice versa (28.8%) are comparable to the in-domain performance of DIM (40.6% on offmychest and 31.3% on happy), suggesting that our CoBERT can generalize well across empathetic conversations in contrasting sentiments.

5.4 Comparison with BERT-adapted Models

To perform a more comprehensive evaluation of CoBERT, we further compare CoBERT with several competitive BERT-adapted models where the sentence encoders are replaced by BERT. We report the results in the middle section of Table 7.

BERT + MemNet (Zhang et al., 2018a): MemNet incorporates persona into context using a Memory Network (Sukhbaatar et al., 2015) with residual connections. The BERT+MemNet model performs slightly worse than Bi-encoder and much worse than our CoBERT, although it achieves slightly faster inference than Bi-encoder.

BERT+DAM (Zhou et al., 2018b): DAM aggregates multi-granularity matching using convolutional layers. The BERT+DAM model performs significantly better than Bi-encoder in R@1, demonstrating the usefulness of learning n-gram matching over the word-word matching matrices. Nevertheless, CoBERT performs noticeably better and has faster inference (7.13x) than BERT+DAM.

BERT+DIM (Gu et al., 2019): The BERT+DIM model combines the benefits from both the strong sentence representation of BERT and the rich finer-

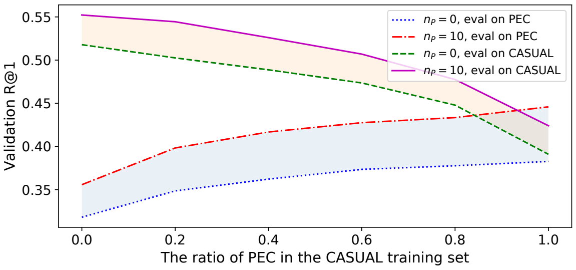

Figure 3: Validation R@1 (in %) against different ratios of PEC in the CASUAL training set.

grained matching of DIM. However, BERT+DIM performs slightly worse than CoBERT, suggesting that the more complex matching and aggregation methods in DIM do not lead to performance improvement over our multi-hop co-attention. In addition, our CoBERT is substantially faster (9.18x) than BERT+DIM in inference, thus more practical in real-world applications.

5.5 Ablation Study

We conduct ablation studies for CoBERT, as reported in the bottom section of Table 7.

Removing either hop-1 or hop-2 co-attention results in noticeably worse performance, albeit slightly faster inference. Removing hop-1 leads to larger performance drop than removing hop-2, suggesting that the first-order matching information seems more important than the second-order matching information for response selection. An additional hop-3 co-attention results in slightly worse performance, suggesting that our two-hop co-attention is the sweet spot for model complexity.

Replacing the max pooling in the hop-1 co-attention by mean pooling leads to much worse performance. In addition, concatenating the results from both max and mean pooling slightly degrades performance, as well as inference speed, suggesting that max pooling may be essential for extracting discriminative matching information.

6 Discussion

6.1 Empathetic vs. Non-empathetic

We investigate whether persona improves empathetic responding more when CoBERT is trained on empathetic conversations than non-empathetic ones. First, we introduce a non-empathetic conversation dataset as the control group, denoted as CASUAL, which is the same as the control group in Section 3 but much larger in size. The CASUAL dataset is collected and processed in the same way as PEC but has significantly lower empathy than PEC (see Table 2). The sizes of training, validation, and testing splits of CASUAL are 150K, 20K, and 20K, respectively. Then, we replace a random subset of training examples from CASUAL by the same number of random training examples from PEC. We then compare the persona improvement, i.e., R@1 ($n_P = 10$) $-$ R@1 ($n_P = 0$), on the PEC validation set and the CASUAL validation set for different replacement ratios.

The results are illustrated in Figure 3. It is unsurprising that for both cases, i.e., $n_P = 0$ and $n_P = 10$, the validation R@1 on PEC increases, and the validation R@1 on CASUAL decreases as the ratio of PEC in the training dataset increases. We also observe that persona consistently improves performance on both validation sets for all ratios.

By investigating the widths of the two shaded regions in Figure 3, we find that the persona improvement on casual responding remains almost constant as more CASUAL training examples are used (3.31% when trained on all 150K PEC conversations vs. 3.44% when trained on all 150K CASUAL conversations). However, the persona improvement on empathetic responding consistently increases as more PEC training examples are used (3.77% when trained on all 150K CASUAL conversations versus 6.32% when trained on all 150K PEC conversations), showing that persona improves empathetic responding significantly more when CoBERT is trained on empathetic conversations than non-empathetic ones ($p < 0.001$, one-tailed t-test).

This result reveals an empirical link between persona and empathy in human conversations and may suggest that persona has a greater impact on empathetic conversations than non-empathetic ones. The result also shows that CoBERT can learn this link during training and use it to perform better empathetic responding during testing. One possible psychological root of this link is that persona is highly correlated to personality (Leary and Allen, 2011), which in turn influences empathy and empathetic responding (Costa et al., 2014). A more detailed analysis of this empirical link is left for future work.

6.2 Number of Persona Sentences

We analyze the persona improvement with respect to different numbers of persona sentences n_P, as

n_P	0	1	2	5	10	20
R@1	40.4	42.0	42.8	45.1	46.2	**47.1**
InfTime	**1.00x**	1.34x	1.38x	1.55x	1.90x	2.96x
RAM	**1.00x**	1.05x	1.06x	1.19x	1.51x	2.29x

Table 8: Validation R@1 (in %), inference time (InfTime) and memory usage (RAM) on PEC against different number of persona sentences n_P.

n_P	seen (57.9%)	unseen (42.1%)	all (100%)
0	40.3	38.5	39.6
10	46.5	43.2	45.1

Table 9: Test R@1 (in %) on PEC against examples with seen or unseen personas. n_P denotes the number of persona sentences.

Context: I'm on a diet and lost ten pounds this month!	
Persona	**Model Response**
I am a college graduate.	Congrats! I'm trying to lose weight as well.
I work in a gym.	Good job! Doing some exercises will help you stay fit!
I am a doctor.	Congrats! Don't forget to take adequate nutrition though.

Table 10: Case study.

shown in Table 8[13]. It is clear that model performance, inference time, and memory usage all increase when more persona sentences are incorporated. Note that memory usage grows quadratically with n_P due to the self-attention operations in BERT. We chose $n_P = 10$ in our experiments because it achieves competitive performance at a reasonable cost of efficiency.

6.3 Performance on New Personas

We analyze the CoBERT performance on examples with new personas. In PEC test set, 42.1% examples are from new speakers. The performance of CoBERT on test examples with seen and unseen (new) speakers is shown in Table 9. The results show that 1) CoBERT performs reasonably well on examples with unseen personas, suggesting that CoBERT can generalize well to unseen personas and retrieve the right response for new speakers accurately; 2) CoBERT performs worse on examples with unseen personas than seen personas; 3) leveraging personas during model training and testing improves CoBERT on examples with either seen or unseen personas; and 4) the persona improvement is more noticeable for examples with seen personas than unseen personas.

6.4 Case Study

We conduct a case study on how persona affects empathetic responding, as shown in Table 10. The model responses are selected by CoBERT from 1K candidates. It is clear that given the same context, different personas lead to different persona-based empathetic responses. For example, when the persona is "I am a doctor.", the model response expresses both praises and caring about the speaker's health.

7 Conclusion

We present a new task and a large-scale multi-domain dataset, PEC, towards persona-based empathetic conversations. We then propose CoBERT, an effective and efficient model that obtains substantially better performance than competitive baselines on PEC, including the state-of-the-art Poly-encoder and several BERT-adapted models. CoBERT is free from hyper-parameter tuning and universally applicable to the task of response selection in any domain. Finally, we present the first empirical study on the impact of persona on empathetic responding. The results reveal an empirical link between persona and empathy in human conversations and may suggest that persona has a greater impact on empathetic conversations than non-empathetic ones.

Acknowledgments

This research is supported, in part, by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI) (Alibaba-NTU-AIR2019B1), Nanyang Technological University, Singapore. This research is also supported, in part, by the National Research Foundation, Prime Minister's Office, Singapore under its AI Singapore Programme (AISG Award No: AISG-GC-2019-003) and under its NRF Investigatorship Programme (NRFI Award No. NRF-NRFI05-2019-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore. This research is also supported, in part, by the Singapore Ministry of Health under its National Innovation Challenge on Active and Confident Ageing (NIC Project No. MOH/NIC/COG04/2017 and MOH/NIC/HAIG03/2017).

[13] Using $n_P = 30$ results in memory error.

References

Scott Brave, Clifford Nass, and Kevin Hutchinson. 2005. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies*, 62(2):161–178.

Zhangming Chan, Juntao Li, Xiaopeng Yang, Xiuying Chen, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Modeling personalization in continuous space for response generation via augmented wasserstein autoencoders. In *EMNLP-IJCNLP*, pages 1931–1940.

Qian Chen and Wen Wang. 2019. Sequential attention-based network for noetic end-to-end response selection. *arXiv preprint arXiv:1901.02609*.

Patricio Costa, Raquel Alves, Isabel Neto, Pedro Marvao, Miguel Portela, and Manuel Joao Costa. 2014. Associations between medical student empathy and personality: a multi-institutional study. *PloS one*, 9(3).

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *ACL*, pages 593–602.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Jiazhan Feng, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Learning a matching model with co-teaching for multi-turn response selection in retrieval-based dialogue systems. In *ACL*, pages 3805–3815.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR Mental Health*, 4(2):e19.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zhu, and Quan Liu. 2019. Dually interactive matching network for personalized response selection in retrieval-based chatbots. In *EMNLP-IJCNLP*, pages 1845–1854.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *ICLR*.

Carl Jung. 2016. *Psychological types*. Taylor & Francis.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jonathan Tarter Klein. 1998. *Computer response to user frustration*. Ph.D. thesis, Massachusetts Institute of Technology.

Mark R Leary and Ashley Batts Allen. 2011. Personality and persona: Personality processes in self-presentation. *Journal of Personality*, 79(6):1191–1218.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *ACL*, pages 994–1003.

Qintong Li, Hongshen Chen, Zhaochun Ren, Zhumin Chen, Zhaopeng Tu, and Jun Ma. 2019. EmpGAN: Multi-resolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019a. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, pages 121–132.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Zihan Liu, and Pascale Fung. 2019b. Caire: An end-to-end empathetic chatbot. *arXiv preprint arXiv:1907.12108*.

K Liu and Rosalind W Picard. 2005. Embedded empathy in continuous, interactive health assessment. In *CHI Workshop on HCI Challenges in Health Assessment*, volume 1, page 3.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, pages 285–294.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *ACL*, pages 5454–5459.

Pierre-Emmanuel Mazare, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *EMNLP*, pages 2775–2779.

Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. 2018. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of Medical Internet Research*, 20(6):e10148.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca

Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, pages 8024–8035.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, pages 5370–5381.

Nadine R Richendoller and James B Weaver III. 1994. Exploring the links between personality and empathic response style. *Personality and Individual Differences*, 17(3):303–311.

Kimberley Rogers, Isabel Dziobek, Jason Hassenstab, Oliver T Wolf, and Antonio Convit. 2007. Who cares? revisiting empathy in asperger syndrome. *Journal of Autism and Developmental Disorders*, 37(4):709–715.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *ACL*, pages 1509–1519.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience lookahead. *arXiv preprint arXiv:1906.08487*.

Farhad Bin Siddique, Onno Kampman, Yang Yang, Anik Dey, and Pascale Fung. 2017. Zara returns: Improved personality induction and adaptation by an empathetic virtual agent. In *ACL*, pages 121–126.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *NIPS*, pages 2440–2448.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Peter Wright and John McCarthy. 2008. Empathy and experience in hci. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 637–646.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *ACL*, pages 496–505.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In *EMNLP-IJCNLP*, pages 111–120.

Rohola Zandie and Mohammad H Mahoor. 2020. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. *arXiv preprint arXiv:2003.02958*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, pages 2204–2213.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018b. Modeling multi-turn conversation with deep utterance aggregation. In *COLING*, pages 3740–3752.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2019. A pre-training based personalized dialogue generation model with persona-sparse data. *arXiv preprint arXiv:1911.04700*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018a. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 0:1–62.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *EMNLP*, pages 372–381.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018b. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*, pages 1118–1127.