Towards Enhancing Health Coaching Dialogue in Low-Resource Settings

Yue Zhou [1], Barbara Di Eugenio[1], Brian Ziebart[1], Lisa Sharp[1], Bing Liu[1], Ben Gerber[2], Nikolaos Agadakos[1] and Shweta Yadav[1]

[1]University of Illinois at Chicago, IL, USA
[2]UMass Chan Medical School, MA, USA
{yzhou232,bdieugen,bziebart,sharpl,liub,nagada2,shwetay}@uic.edu
{ben.gerber}@umassmed.edu

Abstract

Health coaching helps patients identify and accomplish lifestyle-related goals, effectively improving the control of chronic diseases and mitigating mental health conditions. However, health coaching is cost-prohibitive due to its highly personalized and labor-intensive nature. In this paper, we propose to build a dialogue system that converses with the patients, helps them create and accomplish specific goals, and can address their emotions with empathy. However, building such a system is challenging since real-world health coaching datasets are limited and empathy is subtle. Thus, we propose a modularized health coaching dialogue system with simplified NLU and NLG frameworks combined with mechanism-conditioned empathetic response generation. Through automatic and human evaluation, we show that our system generates more empathetic, fluent, and coherent responses and outperforms the state-of-the-art in NLU tasks while requiring less annotation. We view our approach as a key step towards building automated and more accessible health coaching systems.

1 Introduction

Health coaching is a patient-centered, motivational interviewing-based clinical practice that focuses on helping patients identify and accomplish personalized, lifestyle-related goals to improve health behaviors. It has been effective in improving the control of chronic conditions such as diabetes and cardiovascular disease and mitigating mental health conditions such as anxiety and depression (Butterworth et al., 2006; Ghorob, 2013; Kivelä et al., 2014; Thom et al., 2016). Health coaching can be particularly beneficial to low-socioeconomic status (SES) populations who disproportionately suffer physical and mental disease burdens (Thackeray et al., 2004; Kangovi et al., 2014). Yet, it is invariably cost-prohibitive for these populations due to its highly personalized and labor-intensive nature.

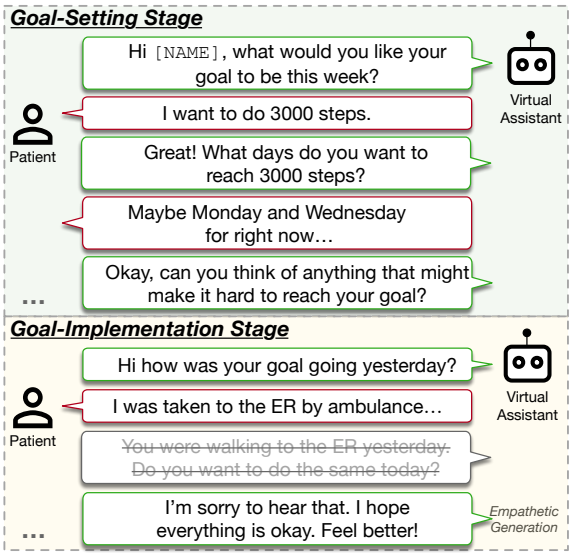

Figure 1: A health coaching dialogue scenario in our dialogue system. It starts with a goal-setting stage where the coach discusses a realistic goal with the patient. After the goal is settled, the coach follows up on the patient's progress and maintains patient engagement. Understanding the patient's emotion cues and responding empathetically is also crucial in such a scenario; the struck-out response, generated by a naïve sequence-to-sequence model, is inappropriate without the capability of understanding and modeling empathy.

Facilitating conversations in healthcare settings between the participants via language techniques and text messages (Aguilera and Muñoz, 2011; Fitzpatrick et al., 2017) has the potential to improve the efficacy of health coaching while reducing the cost. However, the interactions between the dialogue systems and patients are either scripted or with limited natural language understanding (NLU) or generation (NLG) capabilities (Kocielnik et al., 2018; Chaix et al., 2020; Mohan et al., 2020). In our previous work Gupta et al. (2020a,b, 2021), we collected real-world health coaching conversation datasets and focused on the NLU components of the dialogue. However, no existing health coaching dialogue system supports natural language conver-

Proceedings of the 29th International Conference on Computational Linguistics, pages 694–706
October 12–17, 2022.

sations between the patient and a coach agent.

In this paper, we propose to build a dialogue system that converses with patients and helps them create and accomplish specific goals for physical activities based on the health coaching dataset we previously collected (Gupta et al., 2020b). We want the system to emulate the health coaching process, which starts with a goal-setting stage where the coach discusses creating a S.M.A.R.T goal with the patient, namely a goal that is specific, measurable, achievable, relevant, and time-bound (Doran, 1981). Once the goal is settled, the coach agent would follow up on the patient's progress and maintain patient engagement. In addition, we focus on developing the system to understand and address the patient's emotions and respond empathetically, which is crucial for better procedure outcomes in healthcare settings (Levinson et al., 2000; Moudatsou et al., 2020). A health coaching dialogue example is shown in Figure 1.

Building such a dialogue system is challenging: First, the real-world health coaching dataset is limited in size and annotation, which restricts not only end-to-end but also modularized approaches, especially with no dialogue states annotated. Additional annotation is also resource-intensive. Second, generating empathetic responses is subtle, which may require incorporating external empathetic knowledge, since cases of empathetic responses in our health coaching dataset are rare.

To address these challenges, we propose a modularized task-oriented health coaching dialogue system with a simplified architecture that requires fewer annotations. The system contains an NLU module, an NLG_{hc} module, and an NLG_{emp} module. The NLU module aims to keep track of the goal attributes as simplified belief states. The NLG_{hc} module takes as input the current dialogue context, the belief states, and the coaching stage to generate coaching utterances. Finally, for the NLG_{emp} module, we build an emotion cue detector and mechanism-conditioned empathy generator to facilitate empathetic response generation.

To evaluate our approach, we combine automatic evaluation with expert-based human evaluation. Our experimental results show that our NLU module outperforms the state-of-the-art (Gupta et al., 2020a,b, 2021) by > 10% in F1-score in the slot-filling and > 7% in semantic frame correctness in the offline goal attributes tracking task, which also enables updating goal information online at every dialogue turn. Moreover, the experiments demonstrate that our dialogue generation achieved best performance compared to baseline methods in terms of coherence, fluency, and empathy. Finally, through a pilot human evaluation, our model's generation is preferred by the health coaches as concerns coherence and empathy.

The contributions of this work are: (1) We propose to build an efficient modularized health coaching dialogue system that helps patients create and accomplish specific goals, with a simplified NLU and NLG framework combined with mechanism-conditioned empathetic response generation. (2) Our system outperforms the state-of-the-art in NLU tasks while significantly reducing the annotation workload. (3) Through automatic and human evaluation, we show our system generates more coherent and empathetic responses, which can provide suggestions for health coaches and improve coaching efficiency.

2 Related Work

- **Conversational Agents in Healthcare.** Conversational agents have been explored to improve the efficacy and scalability of the interactions between healthcare professionals and patients. For instance, chatbots in different healthcare settings, including chronic disease monitoring (Chaix et al., 2020), cognitive behavior therapy (Fitzpatrick et al., 2017), and physical activity promotion (Mohan et al., 2020; Kocielnik et al., 2018). However, they are limited in natural language understanding and generation capabilities. More sophisticated approaches have been proposed in mental health counseling (Althoff et al., 2016; Shen et al., 2020). In our previous work Gupta et al. (2020a,b, 2021), we collected real-world health coaching conversation datasets and focused on the NLU components of the dialogue which summarize weekly goals to support health coaches.

- **Task-Oriented Dialogue.** Traditional task-oriented dialogue systems are modularized (Jokinen and McTear, 2009). They consist of an NLU component for understanding the user intent and recording the dialogue states and an NLG component for policy management and response generation (Williams et al., 2016; Budzianowski et al., 2018; Mrkšić et al., 2017; Wen et al., 2015). However, approaches

have shifted towards end-to-end architectures to reduce human effort and error propagation between modules (Bordes et al., 2017; Wen et al., 2017). Recently, training an end-to-end system as a sequence prediction problem leveraging the causal language models has delivered promising results (Hosseini-Asl et al., 2020; Peng et al., 2021).

- **Empathetic Data for Conversations.** Empathetic interaction is a key to better task outcomes in conversations. Recently, empathetic data and approaches have been proposed to facilitate empathetic conversations in open-domain and healthcare settings. Babytalk (Hunter et al., 2008; Mahamood and Reiter, 2011) is an earlier system that summarizes neonatal intensive care patient data for different types of users, and provides affective cues for the parents. Rashkin et al. (2019) proposed an open-domain empathetic dialogue dataset (ED), with each dialogue grounded in an emotional context. Welivita et al. (2021) extended from ED and proposed a large-scale silver-standard empathetic dialogue data based on movie scripts. Sharma et al. (2020, 2021a) proposed an empathetic dataset in mental health support settings, with the communication mechanisms and corresponding strength of empathy annotated.

3 Health Coaching Dataset

We first briefly describe our dataset, that motivates the approach that we are proposing; full details can be found in Gupta et al. (2020a). We collected two datasets of health coaching dialogues between patients and coaches via text messages. Dataset 1 and 2 contain 28 and 30 patients, with each patient coached for four and eight weeks, respectively. Each week the health coach converses with the patient to create a physical activity S.M.A.R.T. goal and then follows up on the patient's progress. The two datasets contain 336 weeks of dialogues with 21.4 average turns per week.

We defined ten slots for the goal's attributes (types of activity, amount, time, days, location, duration, and the confidence score for the activity) (the Appendix contains examples of all slots). We used a stage-phase schema for additional turn-level annotation describing how the health coaching dialogue unfolds. We defined two stages: the goal-setting stage and the goal-implementation stage. Each stage includes a set of phases, such as goal identification, negotiation, and follow-up. Each turn can belong to a certain stage-phase combination. Later, we added dialogue act annotations (Gupta et al., 2021), consisting of 12 domain-independent dialogue acts, following the ISO-standard by Mezza et al. (2018).

Since dataset 1 was collected before dataset 2, the manual annotation was developed on dataset 1, and used to develop our NLU component, that was then tested on dataset 2. Hence, dataset 1 is fully annotated for slot-value spans, goals, stages and phases. In dataset 2, only three patients are annotated for slot-value spans and phases, and 15 weeks for goals; the whole dataset is annotated for stages. As far as dialogue acts are concerned, they are only available for 15 weeks of dialogues in dataset 1[1].

4 Methods

In this section, we begin by first providing a brief workflow of our proposed health coaching dialogue system and then describing the model architecture in details.

4.1 Health Coaching Dialogue System

A health coaching dialogue can be framed into stages: **(1)** Starting with the goal-setting stage, the coach helps the patient create a specific goal whose attributes can be represented by a list of slot values[2] (*e.g., Activity = Walk; Amount = 3000 steps; Days = Mon-Fri*), which retains a task-oriented nature. However, it is much more complex than the prevailing task-oriented services, since **(2)** after the goal is initialized, the coach needs to follow up on the patient's progress and maintain patient engagement (e.g., checking in, sending reminders, revising the goal, and providing encouragement on patient): this is the goal implementation stage. A health coaching dialogue contains multiple turns. We denote the stage that the turn t belongs to as S_t.

Leveraging the task-oriented dialogue framework, we use belief state B_t, a list of slot-value pairs, to record the goal attributes in a turn t. An **NLU** module is used to infer B_t by considering the earlier patient utterances $U_{<t}$, current patient utterances U_t, and earlier system responses $R_{<t}$ as

[1]The datasets with annotations are available at https://github.com/uic-nlp-lab/virtualcoachdata.

[2]We provide a complete list of slots and corresponding examples in the Appendix.

input to the module. Formally:

$$B_t = \text{NLU}([U_0, R_1, U_1, R_2, ..., U_t]) \quad (1)$$

B_t summarizes the goal from the dialogue history and will be used for conditional response generation and lexicalizing the generated response.

Then, given the dialogue context C_t in turn t, consisting of the previous two turns [R_{t-1},U_{t-1}], the stage of the previous turn S_{t-1}, and B_t, we build a sequence classifier to predict the stage S_t:

$$S_t = \text{Seq2Label}_{hc}([C_t, B_t, S_{t-1}]) \quad (2)$$

Finally, a delexicalized response R_t is generated via a Seq2Seq neural network given C_t, B_t, and S_t concatenated as a single sequence:

$$R_t = \text{Seq2Seq}_{hc}([C_t, B_t, S_t]) \quad (3)$$

The response can be lexicalized into human readable text using the belief state B_t. Equations 2 and 3 constitute the **NLG$_{\text{hc}}$** module, where we explore replacing the fine-grained action prediction with the coarse stage prediction as proximal dialogue management.

Although **NLG$_{\text{hc}}$** could learn limited empathetic response patterns such as *"sorry to hear that."* from the health coaching data, it does not retain prior empathetic knowledge and explicitly model empathy. However, showing a caring attitude and being empathetic to patients' emotional cues are crucial to patient activation and engagement, leading to better task outcomes. To facilitate the empathetic capability of the system, we build the **NLG$_{\text{emp}}$** module, where the empathetic response $\tilde{R}_t$ is generated conditioned on the patient's previous turn utterance U_{t-1} and communication mechanism signals M:

$$\tilde{R}_t = \text{Seq2Seq}_{emp}([U_{t-1}, M]) \quad (4)$$

We follow Sharma et al. (2020) and consider three communication mechanisms for empathy: *'Emotional Reactions'*, *'Interpretations'*, and *'Explorations'*. Emotional reaction expresses direct emotions (e.g., compassion) to show empathy, such as *"I would be very worried."* Interpretation communicates an understanding of the speaker's experience, such as *"I know anxiety is scary."* Exploration expresses empathy by exploring the speaker's feelings, such as *"What happened? How come?"* An empathetic response can be realized through multiple communication mechanisms. Thus, M contains one or more of the three special tokens `[EMOR]`, `[INTERP]`, and `[EXPLOR]`, representing the three communication mechanisms. We seek to use the mechanism signals to control the style of the generated empathetic response, making it flexible and appropriate for health coaching scenarios.

In addition, an emotion cue detector is built to support empathetic generation. We predict the current emotion signal E_t given the patient's previous utterance U_{t-1}. Then $\tilde{R}_t$ is generated when a strong emotion cue is detected with the probability of certain types of emotion greater than a predefined threshold, *i.e.*, $p(E_t|U_{t-1}) > \tau$, obtained from development set performance.

The overall framework of our model is illustrated in Figure 2.

4.2 Model Architectures

In this section, we describe the detailed model architecture for each of the NLU, NLG$_{\text{hc}}$, and NLG$_{\text{emp}}$ module.

4.2.1 The NLU Module

The NLU module records the current state of the slot values. It consists of a neural slot-filling model that extracts the slot fillers from each utterance and a carryover classifier to determine if the value of a slot should be copied from the previous state.

Neural Slot-Filling A neural slot-filling model maps the sentence representation to a sequence of BIO labels (the beginning *(B)* and inside *(I)* of each slot label, and outside *(O)* for others), often combined with sentence-level classifications (e.g., domain, intent). Following Chen et al. (2019), we use BERT (Devlin et al., 2019) as the model backbone, the `[CLS]` token of the sentence for sentence-level classification, and the final hidden state of the first sub-token of each word at position n for BIO labeling. The network is trained by maximizing the conditional probability:

$$p\left(y^c, \boldsymbol{y^s} \mid \boldsymbol{x}\right) = p\left(y^c \mid \boldsymbol{x}\right) \prod_{n=1}^{N} p\left(y_n^s \mid \boldsymbol{x}\right) \quad (5)$$

Where y^c, y_n^c are softmax probabilities for sentence labels and BIO labels of word n, with $\boldsymbol{x}$ being the sequence of word tokens.

Carryover Classifier The carryover classifier determines if the value of a slot should be copied from the previous state or updated with the new instance seen in the current utterance, which outputs a binary value for each slot.

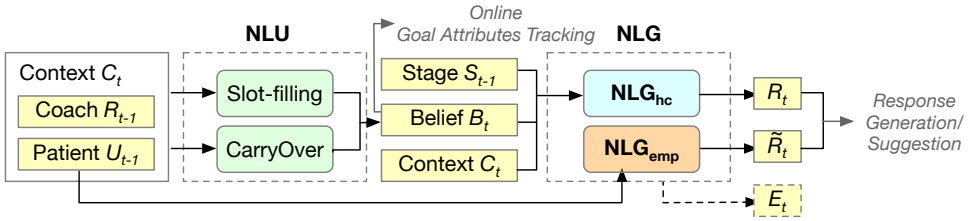

Figure 2: The framework of our health coaching dialogue system. The NLU module consisting of the slot-filling and carryover model reads the dialogue and infers belief state B_t. The NLG_{hc} module takes as input the stage S_{t-1}, belief state B_t, and context C_t to generate response R_t. The NLG_{emp} handles the cases where empathy are required and outputs empathetic response $\tilde{R}_t$ and emotion signal E_t.

Following the work of Gao et al. (2019), we developed our slot carryover classifier based on the current dialogue context. Our proposed carryover classifier, however, differs from Gao et al. (2019), in terms of: (1) we design the classifier as a separate model rather than a component in an end-to-end belief state tracking architecture; (2) it benefits from the results of the slot-filling model and only makes a prediction when there is a collision between the existing value and the new for a given slot. Such a design takes into account that: (1) dialogue datasets in healthcare are limited in size, making end-to-end training difficult; (2) these datasets can also be limited in annotations. Our model can facilitate annotation since the annotators can work with the slot-filling model to only examine the lines where the value collision occurs, enabling more efficient labeling for belief states in a carryover fashion.

We use BERT to encode the contexts and utilize the hidden state associated with the `[CLS]` token to represent the current context C_t and maximize $p(y_t|C_t)$ via fine-tuning BERT with cross entropy minimization:

$$h_t = \text{BERT}(C_t)[\text{CLS}]$$
$$y_t = \text{Softmax}(Wh_t + b)$$

where $y_t \in [0, 1]^{N_s}$, N_s is the number of slots.

During inference, the slot-filling model extracts the slot fillers at each turn. The carryover classifier determines if the value should be copied from the previous state or updated with the new instance given dialogue context when there is a value conflict. NLU enables an update for the belief state B_t at each turn, *i.e.,* providing online goal attributes tracking.

4.2.2 The NLG_{hc} Module

The dominant modularized approach for task-oriented dialogue often consists of belief state tracking, dialogue policy, and language generation. It requires dedicated modeling and annotations for dialogue acts. However, annotating acts is resource-intensive in healthcare settings.

An alternative is using stages, where each stage restricts a set of possible actions in health coaching dialogue and other healthcare conversations, such as clinical motivational interviews and patient encounters. For example, in health coaching, sending reminders and encouragement on progress can only appear in the goal implementation stage when the goal has already been created. Another example is the SOAP (Subjective, Objective, Assessment, and Plan) structure (Podder et al., 2021) of patient encounters, in which discussing a patient's chief complaint occurs in the Subjective part, while diagnosis is discussed in the Assessment part.

Thus, we explore the possibility of using stages that contain coarse and fuzzy act information to guide dialogue generation instead of fine-grained act annotations. Concretely, we use T5 (Raffel et al., 2020) to jointly model Seq2Label_{hc} (*cf.*2) and Seq2Seq_{hc} (*cf.*3) for stage S_t prediction and response R_t generation as a multi-task approach. The input is a single sequence by concatenating the local dialogue context, current belief state, and stage tokens. For example, the input for Seq2Seq_{hc} is the concatenation of C_t, B_t, and S_t, separated by delimiter tokens, which is mapped to the target response sequence, as shown in Figure 3. The model is encoder-decoder based, trained with a maximum likelihood objective with a different prefix prepending to the input corresponding to each task.

4.2.3 The NLG_{emp} Module

Empathetic Response Generation We fine-tune GPT2 (Radford et al., 2019) for empathetic response generation. At training time, we concatenate the user utterance, the corresponding empathetic response, and the communication mecha-

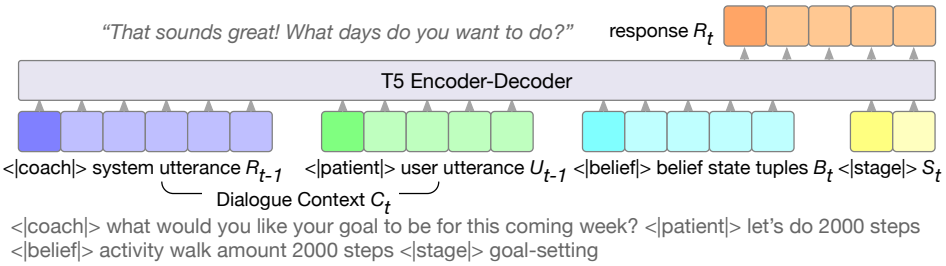

Figure 3: Model architecture of generating R_t in NLG_{hc}. The context C_t, belief tokens B_t, and the predicted stage S_t are concatenated as a single input sequence, training with a T5 encoder-decoder model.

nisms used in the response, separated by delimiter tokens, as one single training sequence $\boldsymbol{x}$. We want GPT2 to learn to model the joint probability $p(\boldsymbol{x})$ by fine-tuning on empathetic data. During inference time, the model takes as input the user utterance and the communication mechanisms and generates one token at a time for empathetic response.

The following shows an example of such a training sequence:

> <|bos|> [EMOR] I was so exhausted yesterday. <|sep|> That's understandable. Take some rest! <|eos|>

where <|bos|> and <|eos|> are the delimiters for beginning of sequence and end of sequence, [EMOR] stands for the communication mechanism *Emotional Reaction*, and <|sep|> separates the user utterance and empathetic response . During inference, the model starts to generate tokens after <|sep|>.

Emotional Cue Detection To support empathetic response, we build a BERT-based multi-class classifier to predict the emotion signals of the patient's utterance. In this work, we use the softmax probability for the predicted label to determine if the empathetic response needs to be generated. However, the predicted emotion labels also facilitate future sentiment analysis work in health coaching dialogue.

5 Experiments

This section includes datasets description, evaluation metrics, experimental results, and qualitative analysis.

5.1 Dataset Details

We conducted our experiments on the following benchmark datasets:

Health Coaching Datasets These are the datasets we described earlier in Section 3. Following Gupta et al. (2020b), we use dataset 1 for training/development and dataset 2 for testing for both NLU and NLG_{hc} to ensure comparability.

Data Augmentation for Slot-filling We only consider the utterances that contain at least one slot-value span for slot-filling. This results in 955:105:205 data points for training, development, and testing. To alleviate data scarcity, we use a naïve data augmentation method to synthesize training examples that (1) first randomly replaces the value with one possible alternative for each slot; (2) then rephrases the modified sentence with a pre-trained paraphrase model[3].

Empathetic Datasets We utilized two benchmark empathetic dialogue datasets: *(i)* EMPATHETICDIALOGUES (ED) (Rashkin et al., 2019) and *(ii)* EPITOME. EMPATHETICDIALOGUES (ED) is a open-domain empathetic dialogue dataset containing 24,850 dialogues, with each dialogue grounded in one of 32 emotional context (e.g., proud, apprehensive, confident, guilty). EPITOME (Sharma et al., 2020) is an empathetic dataset in mental health support settings. The dataset consists of 3k posts annotated with respect to the communication mechanisms (*Emotional Reaction, Exploration, Interpretation*) and level of empathy (from 0 to 2, representing no empathy, weak and strong empathy, respectively) in each mechanism.

To facilitate empathetic generation conditioned on the communication mechanisms, we first build a weak multi-label classifier for predicting the types of communication mechanisms based on the 3k posts in Sharma et al. (2020) (achieving a hamming loss (Tsoumakas and Katakis, 2007) of ∼0.16). Then we apply the classifier on the ED dataset

[3] A fine-tuned PEGASUS (Zhang et al., 2020) on PAWS (Zhang et al., 2019)

to get communication mechanism labels for each instance. We use their default train/dev/test split. Finally, $N = 64$ empathetic cases in the health coaching dataset are used for few-shot fine-tuning the trained empathetic generation model.

5.2 Evaluation Metrics

Depending upon the task, we use the following evaluation metrics:

1. Slot-filling: Precision, Recall, and F1-score.
2. Dialogue State (Goal Attribute) Tracking: partial/complete match and goal correctness@k following Gupta et al. (2020b):

   Correctness@k: Computes the percentage of correctness over all the predicted goals of each week. If at least k attributes are predicted correctly, the goal is regarded as correct.
3. Dialogue Generation: BLEU (Papineni et al., 2002), BertScore (Zhang et al., 2020), Perplexity (PPL), and Empathy Score.

   Perplexity: We measure fluency as perplexity (PPL) of the generated response using a pre-trained GPT2 model that has not been fine-tuned for this task, following previous work (Ma et al., 2020; Sharma et al., 2021b).

   Empathy Score: We train a standard text regression model based on BERT using the response posts and corresponding level of empathy scores in Sharma et al. (2020) (achieving an RMSE of $\sim$0.57). We use this model to measure the empathy in the generated outputs.

For detailed descriptions of metrics, training, including model parameters, selection, and supplementary analysis, please see the Appendix.

System	Slot R	Slot P	Slot F1	PAcc
Gupta et al. (2020b)	0.806	0.808	0.790	0.801
+Phase	0.899	0.837	0.867	0.779
+StartPhase	0.910	0.847	0.877	**0.835**
+StartPhase+Aug	**0.926**	**0.879**	**0.902**	0.817
Slot Only+Aug	0.904	0.876	0.890	-

Table 1: Evaluation on slot-filling with ablations. *Aug*: using data augmentation; *StartPhase*: Jointly predicting the phase of the sentence only if it is the beginning sentence of the phase.

5.3 Results

5.3.1 Goal Attributes Tracking

Table 1 shows the performance of slot-filling and phase prediction compared to the previous model. Jointly predicting the phase of the sentence only if it is at the beginning of the phase, combined with data augmentation (+StartPhase+Aug), achieved the best performance for slot-filling, outperforming the state-of-the-art by 11.2% in F1. However, modeling without phases suffices for slot-filling while reducing the annotation cost. As such, we adopt the no-joint model for downstream tasks. Our experimental result also shows the carryover classifier achieved a F1-score of 0.88 using only the dialogue context. We investigate whether dialogue act and phase labels can improve carryover classifier, however they barely contribute to the model performance[4].

In previous work, we extracted goals at two critical points for each week to evaluate offline goal tracking: one at the end of the goal-setting stage (forward) and the other at the end of the goal-implementation stage (backward). The forward and backward goals can be different since the patient may encounter barriers, and the goal can be revised in the implementation stage. We also proposed a rule-based approach to update the slot values, which simply records the last mention of the value for each slot except for certain conditions. In this paper, we compared our model with previous work and a combination of our slot-filling model with the previous rule-based approach (our SF+Rule). Table 2 shows the performance for goal attributes tracking of our NLU module compared with previous work. For dataset 1 backward goals, our SF+Rule achieved the best performance resulting from more accurate slot-filling. We observe that in dataset 1, the coach tends to summarize the goal to the patients at the end of each week, which benefits the rule-based approach. Nonetheless, our NLU module outperforms previous work in all evaluation metrics on dataset 1 forward and dataset 2; it also enables goal tracking at each turn (online).

5.3.2 Dialogue Generation

Since our previous work did not include a generation model, we compare our dialogue generation with three baselines: (1) Retrieval-based model. We use BERT to encode the current belief state and dialogue context as query h_q, and encode the

[4]See Appendix for detailed results.

	Dataset 1 Backward					Dataset 1 Forward					Dataset 2				
	Complete	Partial	@k=10	@k=9	@k=8	Complete	Partial	@k=10	@k=9	@k=8	Complete	Partial	@k=10	@k=9	@k=8
Gupta et al. (2020b)	0.77	0.81	13.1	39.3	**65.4**	0.81	0.85	15.0	44.9	68.2	-	-	-	-	-
ourSF+Rule	**0.808**	**0.858**	**31.6**	**42.1**	63.2	0.831	0.863	21.7	69.6	73.9	0.767	0.787	13.6	18.2	40.9
ourNLU	0.792	0.825	15.8	36.8	63.2	**0.838**	**0.869**	**21.7**	**69.6**	**78.3**	**0.8**	**0.813**	**13.6**	**27.3**	**63.6**

Table 2: Evaluation on offline goal attributes tracking with complete/partial match and goal correctness@k.

Input	**Model**	**Generated Response**
Coach: What would you like your goal to be this week?	Ours	Sounds good, which days were in your mind?
Patient: I want to walk 30 min a day between 6am to 8am.	IR	And what days would you like to walk the track around neighborhood?
	Seq2Seq	How many days do you want to reach this week?
Coach: Good morning! How is your goal for this week going?	Ours	`[EMOR]` Oh no, I hope you are okay.
Patient: I'm sorry I didn't go to work today I have a massive migraine headache.		`[EXPLOR]` Oh geez, sorry to hear that. Are you feeling better?
		`[INTERP]` I've had this experience before. Sometimes it really hits you.
	IR	Check your steps.
	Seq2Seq	Have you been busy with the migraine this week and how did it turn out?
Coach: Don't forget to take a deep breath - breathing is important.	Ours	Don't forget to take a deep breath - breathing is important.
Patient: You made me smile. Thanks.	✗ Repeat	
Coach: Hi when you get a chance can you sync your fitbit? I only see your steps since last Friday. Thanks.	Ours	That's great news. I hope you don't get into a long trip there.
Patient: Sorry been in Texas, just made it back to Chicago.	✗ Focus	

Table 3: Qualitative examples of generated responses in empathetic and non-empathetic scenarios, combined with error analysis.

Model	BLEU (↑)	PPL (↓)	BertS F1 (↑)	EmpS (↑)
IR	0.194	24.2	0.853	-0.179
Seq2Seq	0.242	16.26	0.863	+0.073
+Acts	0.235	16.2	0.861	-0.065
OURS	**0.251**	**15.6**	**0.872**	**+0.256**

Table 4: Evaluation on dialogue generation with automatic metrics. BLEU: Average of BLEU-1,-2,-3,-4. EmpS: Computed by the difference of the empathetic score between the output and the ground truth ($\sim$ 0.978).

response as h_r. We fine tune the BERT models to retrieve the response that maximizes $h_q \cdot h_r$; (2) A seq2seq neural network that maps the belief state and dialogue context to the output without stages. (3) using dialogue acts instead of stages.

Table 4 shows that our response generation outperforms all baselines in all metrics. The predicted stage information (achieved an accuracy of $\sim$0.92) provides meaningful signals for dialogue generation. In addition, by incorporating external empathetic knowledge, our model achieved +0.256 average improvement in empathy. Including dialogue acts does not improve the performance compared to the seq2seq model. This is not unexpected because labels are imbalanced and only available for 15 weeks of data. Computationally labeling dialogue acts also leads to large error propagation. The emotional cue detection (32-category classification) achieved an accuracy of $\sim$0.58. Based on development set observation, we set τ to be 0.7 and generate empathetic response when the cumulative probability of the top-2 predicted emotional labels is greater than τ.

Human Evaluation. We performed an expert-based human evaluation on coherence and empathy through A/B testing. We ask two health coaches to compare outputs from our model against other baselines given the same input and choose (a) the response which is more empathetic; (b) the response which is more coherent. Among the collected 43 examples, our model's outputs have a $\sim$71% preference for empathy and $\sim$55% preference for coherence over other baselines.

5.4 Qualitative Analysis

We present examples of our health coaching dialogue generation and mechanism-conditioned empathetic generation in Table 3. In the first case, where empathy is not required, all three models choose to inquiry for more information of the goal. However, the fixed, retrieval-based response can contain context-irrelevant tokens (*"..the track around neighborhood"*) and the Seq2Seq response is relatively unnatural. In the second case, given the cues including *"..a massive migraine headache.."*, our model can generate corresponding empathetic responses given different communication mechanisms (e.g., `[EMOR]` $\rightarrow$ *"Oh no, I hope you are okay."*). In contrast, the retrieval-based and Seq2Seq model failed to response empathetically. Particularly, the Seq2Seq model failed to distinguish a symptom (*i.e., "migraine headache"*) from a physical activity or goal. Finally, we present two incoherent generation examples by our model. In the first example, our system misinterprets the

coach's utterance - a *parody* of a reminder - as a real reminder, which tends to repeat again; thus, the system naïvely copies it. In the second example, the system needs to see more context to understand that the trip is an explanation from the patient for not making progress, which the coach asked about in the previous utterance. A better response should address the patient's explanation while maintaining specificity on the trip scenario, e.g., *"No problem. Welcome back."*

6 Conclusions and Future Work

We built an efficient health coaching dialogue system that helps patients create and accomplish specific goals, with a simplified NLU and NLG framework combined with mechanism-conditioned empathetic response generation. The experiments show that the system can generate more coherent and empathetic responses, supporting health coaches and improving coaching efficiency. In addition, our system outperforms the state-of-the-art in NLU tasks while requiring fewer annotations. We view our approach as a key step towards building automated and more accessible health coaching systems in low-resource settings and believe our approach may also generalize to building dialogue systems in similar scenarios, such as patient education at discharge or consulting on behavior change problems.

In the future, we will explore the following directions: (1) Modelling empathetic understanding and generation with the goal response generator as one integrated end-to-end system while providing explainability. (2) A more comprehensive human evaluation from both the coach's and the patient's perspectives, including but not subject to goal completion and activity engagement rate.

Acknowledgments

This work is supported by the National Science Foundation under Grant IIS-1838770.

References

Adrian Aguilera and Ricardo F Muñoz. 2011. Text messaging as an adjunct to cbt in low-income populations: A usability and feasibility pilot study. *Professional Psychology: Research and Practice*, 42(6):472.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Susan Butterworth, Ariel Linden, Wende McClay, and Michael C Leo. 2006. Effect of motivational interviewing-based health coaching on employees' physical and mental health status. *Journal of occupational health psychology*, 11(4):358.

Benjamin Chaix, Arthur Guillemassé, Pierre Nectoux, Guillaume Delamon, Benoît Brouard, et al. 2020. Vik: a chatbot to support patients with chronic diseases. *Health*, 12(07):804.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint arXiv:1902.10909*. Only available as unpublished pre-preprints.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

George T Doran. 1981. There's a SMART way to write management's goals and objectives. *Management review*, 70(11):35–36.

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):e7785.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 264–273, Stockholm, Sweden. Association for Computational Linguistics.

Amireh Ghorob. 2013. Supplement: Health coaching: Teaching patients to fish. *Family practice management*, 20(3):40–42.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020a. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256, 1st virtual meeting. Association for Computational Linguistics.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Bing Liu, Ben Gerber, and Lisa Sharp. 2020b. Goal summarization for human-human health coaching dialogues. In *The Thirty-Third International Flairs Conference*.

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2021. Summarizing behavioral change goals from sms exchanges to support health coaches. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–289.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Jim Hunter, Albert Gatt, Francois Portet, Ehud Baruch Reiter, and Gowri Somayajulu Sripada. 2008. Using natural language generation technology to improve information flows in intensive care units. In *18th European Conference on Artificial Intelligence (ECAI 2008)*, pages 678–682. IOS Press.

Kristiina Jokinen and Michael McTear. 2009. Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies*, 2(1):1–151.

Shreya Kangovi, Frances K Barg, Tamala Carter, Kathryn Levy, Jeffrey Sellman, Judith A Long, and David Grande. 2014. Challenges faced by patients with low socioeconomic status during the post-hospital transition. *Journal of general internal medicine*, 29(2):283–289.

Kirsi Kivelä, Satu Elo, Helvi Kyngäs, and Maria Kääriäinen. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient education and counseling*, 97(2):147–157.

Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection companion: a conversational system for engaging users in reflection on physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–26.

Wendy Levinson, Rita Gorawara-Bhat, and Jennifer Lamb. 2000. A study of patient clues and physician responses in primary care and surgical settings. *Jama*, 284(8):1021–1027.

Xinyao Ma, Maarten Sap, Hannah Rashkin, and Yejin Choi. 2020. PowerTransformer: Unsupervised controllable revision for biased language correction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7426–7441, Online. Association for Computational Linguistics.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 12–21.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. ISO-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Shiwali Mohan, Anusha Venkatakrishnan, and Andrea L Hartzler. 2020. Designing an ai health coach and studying its utility in promoting regular aerobic exercise. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(2):1–30.

Maria Moudatsou, Areti Stavropoulou, Anastas Philalithis, and Sofia Koukouli. 2020. The role of empathy in health and social care professionals. In *Healthcare*, volume 8, page 26. Multidisciplinary Digital Publishing Institute.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2021. Soap notes. In *StatPearls [Internet]*. StatPearls Publishing.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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 Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021a. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA. Association for Computing Machinery.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021b. *Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach*, page 194–205. Association for Computing Machinery, New York, NY, USA.

Ashish Sharma, Adam Miner, David 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)*, pages 5263–5276, Online. Association for Computational Linguistics.

Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. 2020. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 1st virtual meeting. Association for Computational Linguistics.

Rosemary Thackeray, Ray M Merrill, and Brad L Neiger. 2004. Disparities in diabetes management practice between racial and ethnic groups in the united states. *The Diabetes Educator*, 30(4):665–675.

David H Thom, Jessica Wolf, Heather Gardner, Denise DeVore, Michael Lin, Andy Ma, Ana Ibarra-Castro, and George Saba. 2016. A qualitative study of how health coaches support patients in making health-related decisions and behavioral changes. *The Annals of Family Medicine*, 14(6):509–516.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Jason D Williams, Antoine Raux, and Matthew Henderson. 2016. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proc. of NAACL*.

A Training Details

All the following models use Huggingface Transformers Library (Wolf et al., 2020). The hyperparameters are not extensively fine-tuned.

- **Slot-filling.** We use BERT-base as model backbone and associated tokenizer, with max sequence length of the tokenized input set to 50. The model was trained for {**5.0**, 7.0, 10.0}

epochs by Adam with a learning rate of {2e-5, **5e-5**}, batch size of {**32**, 64}.

- **Carryover Classifier.** We use BERT-base as model backbone and associated tokenizer, with max sequence length of the tokenized input set to 96. The model was trained for {5.0, **7.0**, 10.0} epochs by Adam with a learning rate of {2e-5, **5e-5**}, batch size of {**16**, 32, 64}.

- **NLG$_{\text{hc}}$** We use T5-base as model backbone and associated tokenizer, with max sequence length set to 128. The model was trained for 10.0 epochs by AdamW with a learning rate of 1e-4, warm up steps = 400, batch size of 64. We use sampling during decoding with top-k set to 50, top-p set to 0.95.

- **Empathetic Generation** We use GPT2 as model backbone and associated tokenizer, with max sequence length set to 96. The model was first trained for 10.0 epochs on the ED dataset by Adam with a learning rate of 1e-4, warm up steps = 400, batch size of 32. We use sampling during decoding with top-k set to 50, top-p set to 0.95. Then, the model was fine-tuned on 64 examples of health coaching empathetic data by 1 epoch. When decoding at inference, we use sampling with top-k set to 50, top-p set to 0.95.

- **Emotion Detection** We use BERT-base as model backbone and associated tokenizer, with max sequence length set to 96. The model was trained for 8.0 epochs by Adam with a learning rate of 4e-5, and batch size of 32.

B Health Coaching Dataset Slot Examples

Table 5 shows the description of the ten slots with value examples.

C Evaluation Metrics Description

Partial/Complete Match If all the values are correctly recorded for a given slot, it is considered a complete match. If the values are partially correct, it is considered a partial match.

Goal Correctness@k Computes the percentage of correctness over all the predicted goals of each week. A goal contains ten attributes (slot-values); if at least k attributes are predicted correctly, the goal is regarded as correct. Goal Correctness@k is trivially equal to 100% when $k = 0$.

BLEU BLEU score (Papineni et al., 2002) measures the word-level overlap between the generated output and the gold reference response.

BertScore BertScore (Tianyi Zhang et al., 2020) measures the semantic similarity between the generated output and the reference leveraging BERT contextual embeddings.

Fluency We measure fluency as perplexity (PPL) of the generated response using a pre-trained GPT2 model that has not been fine-tuned for this task, following previous work (Ma et al., 2020; Sharma et al., 2021b).

Empathy We train a standard text regression model based on BERT using the response posts and corresponding level of empathy scores in Sharma et al. (2020) (achieving an RMSE of $\sim$0.57). We use this model to evaluate the empathy in the generated output compared with baselines.

D Supplementary Analysis

D.1 Carryover Ablation

Table 7 shows the model performance of the carryover classifier with ablations. Using a combination of phases and acts can slightly improve recall with a tradeoff of reducing precision. However, using the dialogue context alone suffices for carryover classification with less annotation cost.

D.2 Emotion Cue Detection

Table 6 shows the predicted emotion labels with corresponding probabilities by emotion cue detection on some patient's utterance examples. The model can reasonably detect emotion signals of the patient's utterance only trained on the ED dataset.

Slot	Description	Value Examples
[activity]	The type of activity the patient will perform.	*"walk, jogging, stair climbing"*
[amount]	The quantity of activity.	*"2000 steps, 6 fligths"*
[duration]	The duration of the activity.	*"20 min, half an hour"*
[distance]	The distance of the activity.	*"3 blocks, 2 miles, from home to bus stop"*
[time]	The time of the day for the activity.	*"at noon, after lunch, 4 pm"*
[location]	The location of the activity.	*"at work, at home, around the park"*
[dayname]	The days to do the activity.	*"Monday, Tuesday"*
[daynumber]	The number of days for the activity.	*"3 days, 5 days"*
[repeatation]	The frequency of activity.	*"twice a day, daily"*
[score]	The confidence or attainability score.	Range from [1,10]

Table 5: The slot-value schema used in Gupta et al. (2020a).

Example Utterances	Top-2 Predicted Emotion Labels
Sorry I left my fitbit in the emergency room yesterday.	Guilty (0.506), Ashamed (0.323)
I'm not feeling very well yesterday so I did not go out for a walk.	Disappointed (0.305), Ashamed (0.174)
I reached 10k steps last week can you believe that?	Surprised (0.548), Proud (0.229)
Ok.	Angry (0.127), Furious (0.059)
I want to walk 3000 steps today.	Hopeful (0.484), Confident (0.132)

Table 6: The predicted emotion labels with corresponding probabilities by emotion cue detection on the patient's utterances.

Input	P	R	F1	Acc
Context Only	**0.91**	0.86	**0.89**	0.88
Context+Phase	0.88	0.87	0.88	0.87
Context+Act	0.88	0.86	0.87	0.87
Context+Phase+Act	0.90	**0.87**	**0.89**	**0.89**

Table 7: Model performance of carryover classification with ablations.