Human-Human Health Coaching via Text Messages: Corpus, Annotation, and Analysis

Itika Gupta,[1] **Barbara Di Eugenio,**[1] **Brian Ziebart,**[1] **Aiswarya Baiju,**[1] **Bing Liu,**[1]
Ben S. Gerber,[2] **Lisa K. Sharp,**[3] **Nadia Nabulsi,**[3] **Mary H. Smart**[3]
[1]Department of Computer Science
[2]Department of Medicine
[3]Department of Pharmacy Systems, Outcomes, and Policy
University of Illinois at Chicago, Chicago, Illinois
{igupta5, bdieugen, bziebart, abaiju2, liub}@uic.edu
{bgerber, sharpl, nnabul2, msmart5}@uic.edu

Abstract

Our goal is to develop and deploy a virtual assistant health coach that can help patients set realistic physical activity goals and live a more active lifestyle. Since there is no publicly shared dataset of health coaching dialogues, the first phase of our research focused on data collection. We hired a certified health coach and 28 patients to collect the first round of human-human health coaching interaction which took place via text messages. This resulted in 2853 messages. The data collection phase was followed by conversation analysis to gain insight into the way information exchange takes place between a health coach and a patient. This was formalized using two annotation schemas: one that focuses on the goals the patient is setting and another that models the higher-level structure of the interactions. In this paper, we discuss these schemas and briefly talk about their application for automatically extracting activity goals and annotating the second round of data, collected with different health coaches and patients. Given the resource-intensive nature of data annotation, successfully annotating a new dataset automatically is key to answer the need for high quality, large datasets.

1 Introduction

A sedentary lifestyle significantly increases the risk of numerous diseases such as type 2 diabetes, cardiovascular disease, and depression (Booth et al., 2017). Unfortunately, physical inactivity has progressively increased over the past several decades. It can be attributed to using modes of transportation for short distances, labor-saving devices, and less active occupations among various other reasons. However, the underlying problem is a lack of motivation. Successfully implementing healthy behaviors require significant motivation that most people, individually, find difficult to initiate and maintain (Cerin et al., 2010; Poncela-Casasnovas et al., 2015). Health coaching (HC) has been identified as a successful method for facilitating health behavior changes by having a professional provide evidence-based interventions, support for setting realistic goals, and encouragement for goal adherence (Kivelä et al., 2014). But HC has its limitations such as it is expensive, time-intensive, and not available around the clock.

Therefore, we aim to build a dialogue-based virtual assistant health coach that will converse with the patients via text messages and help them to set Specific, Measurable, Attainable, Realistic and Time-bound (S.M.A.R.T.) goals (Doran, 1981). The SMART goal-setting approach has been rigorously adopted to set realistic and manageable goals in different fields such as health behavior change and software engineering. It has been shown that goal setting and action planning help patients adopt healthy behaviors and manage chronic diseases (Bodenheimer et al., 2007; Handley et al., 2006). Also, text messages have been shown to help patients follow healthy behaviors as they provide a continuous means of education, support, and motivation (Chow et al., 2015); currently, a majority of the population owns a cellphone (96% are cellphone users and 81% are smartphone users[1]).

Most goal-oriented dialogue systems assume that a user has a predefined goal that needs to be accomplished using the system. However, that is not the case in HC dialogues. Instead of the usual information-seeking dialogues, where the user requests information from the system, HC dialogues are collaborative where both the coach and the patient negotiate a goal that best suits the patient's lifestyle and seems realistic based on their previous activity patterns. An excerpt from dataset 1 is shown in Figure 1. The patient starts with an

[1]https://www.pewresearch.org/internet/fact-sheet/mobile/

(1) Patient: Good morning, my goal is to aim for 30,000 steps, 40 flights, 12000 calories and 12 miles by the end of Friday this week
(2) Coach: Wow that's a lot.
(3) Coach: Last week you did 38 flights. Do you know how many step you got?
(4) Coach: Ok i just calculated roughly 23k for Mon to fri. You had more over the weekend.
(5) Coach: Those are great personal goals to have. Let's focus on the walking goals for the purpose of the study. So how likely do you think you will be able to accomplish your goal of 30K steps and 40 flights?
(6) Patient: Well considering it will be measured Monday through Friday I guess I should reduce my goals.ill aim for 20,000 steps and 30 flights. I feel I will be more
(7) Patient: likely to accomplish this goal without any problems.

Figure 1: Example of a conversation between the health coach and a patient

ambitious goal in (1), and the coach helps to make it more realistic through (2)-(5). As the conversation takes place over the text messages, a dialogue system will also need to take care of abbreviations and typing errors such as 'goals.ill' in (6).

Moreover, most existing dialogue datasets do not involve any follow-up conversations. For instance, once a flight is booked, the system doesn't follow-up on how the trip was or if the user would like to modify the booking. However, it is a crucial step in HC conversations as patients tend to change their goals on encountering a barrier. Lastly, HC conversations happen over multiple days. Some days no messages are exchanged and some days more than 10 messages get exchanged. Most publicly available datasets assume that the task will be finished in one sitting. Due to this collaborative negotiation setting over multiple days in our corpus, goal information is spread throughout the dialogue.

Motivated by these complexities, we decided to annotate our data for two types of information: (1) the SMART goal attributes in the dialogues to track patients' goals, and (2) different stages and phases that model the conversation flow in HC dialogues. For our domain, SMART goal attributes are the slot-values pertaining to a patient's goal. Stages and phases are more abstract, but otherwise analogous to tasks and sub-tasks as defined in task-oriented dialogue systems (Chotimongkol and Rudnicky, 2008). We believe the SMART annotation schema that we designed can be applied to any task where SMART goal setting is being used and not just physical activity. Similarly, the stages-phases annotation schema can be used to model the flow of any collaborative decision making counseling dialogue. In this paper, we will discuss the two rounds of data collection process, the subsequent analysis of the dialogues, which includes developing schemas and annotating the data, and application of models trained on these annotations.

Our contributions can be summarized as follows:

- We describe the data collection methodology for health coaching dialogues via text messages that span over multiple days. We undertook two rounds of data collection; we discuss what we learned in round 1 and what this led us to change in round 2. We will refer to the first round of data as *dataset 1* and the second round of data as *dataset 2* throughout the paper.
- We believe we are the first to formalize the SMART goal-setting approach, which we did based on dataset 1 using two annotation schemas. We demonstrate that this approach results in reliable annotator agreement.
- We show that supervised classification models trained on dataset 1 can be used to automatically extract goals and reliably annotate dataset 2 for SMART tags (macro F-score = 0.81) even though the latter was collected with 3 different health coaches and 30 different patients.
- We will release dataset 2 to the community, since we collected consent from the patients in this regard[2]. Dataset 2 will be available upon request along with the annotation manual. Given the nature of the dataset, out of an abundance of respect for our patients, the text data will not be made public online.

2 Related Work

One cannot build a good domain-specific dialogue system without having any insights into how users will interact with the system. Therefore, first we need data that represents at least some range of actions that are found in human-human or human-machine conversations in the given domain. Initiatives such as the Dialogue State Tracking Challenge (DSTC) started in 2013 to provide a common testbed for different tasks related to domain-specific dialogue systems such as dialogue state

[2]Unfortunately, the activity data collected via Fitbit cannot be shared, since consent did not include permission for such data; dataset 1 cannot be shared, because of lack of consent.

tracking, dialogue act prediction, and response generation; labeled datasets for each of these tasks were provided (Williams et al., 2013). However, most of these datasets focused on traveling and restaurant booking domains (Henderson et al., 2014). Moreover, for data collection, predefined scenarios are given to the users and thus, the users' responses are not as spontaneous as they would be in a real-life situation (Asri et al., 2017; Budzianowski et al., 2018). Unfortunately, there are no such publicly available datasets for dialogue systems in the health domain.

The idea of automated conversational agents to promote healthy behaviors has recently gained considerable interest. Researchers such as Watson et al. (2012) and Both et al. (2010) respectively worked on promoting physical activity adherence and supporting psychotherapy for adults using automated systems. But internally most of these systems rely on a predefined set of input/output mappings, focus more on general goal setting, and do not provide follow-up during goal accomplishment.

Researchers have also focused on computational analysis of conversations in the health domain. Pérez-Rosas et al. (2018) collected Motivational Interviewing (MI) based counseling interviews from public sources such as YouTube and built models to predict the overall counseling quality using linguistic features. Before the YouTube data, the authors also worked on data collected in clinical settings, graduate student training and such, but didn't release it due to privacy reasons (Pérez-Rosas et al., 2016). The authors used the well established Motivational Interviewing Treatment Integrity (MITI) coding system to annotate the data and score how well or poorly a clinician used MI (Moyers et al., 2016). The MITI coding system was also used by Guntakandla and Nielsen (2018) to annotate reflections in the health behavior change therapy conversations. Since MI based interventions focus on understanding patient's attitudes towards the problem and persuading them to change, the MITI coding system supports assessing clinicians based on how well they bring forth patient's experiences, cultivate change talk, provide education, persuade them through logical arguments, and such. However, specific goal setting is not the main focus of these interviews and is rarely discussed.

A framework for health counseling dialogue systems closest to ours is by Bickmore et al. (2011). Their task model comprises opening, small talk, review tasks, assess, counseling, assign task, preclosing, and closing. Conversely, our stages-phases schema looks at the fine-grained decomposition of review-tasks, counseling, and assign task, which Bickmore et al. (2011) did not do. As far as we know, no other work models HC dialogues collected in a SMART goal setting, focusing on slot-values and higher-level conversation flow.

3 SMART Goal Setting

Based on the domain, practitioners modify the definition of SMART components to fit the task at hand. For physical activity, we define them as follows:

- **Specific (S)**: Create a clear goal that is as specific as possible and focuses on a particular activity or task such as cycling, walking, or taking stairs.
- **Measurable (M)**: Quantify the goal to know when the goal has been accomplished.
- **Attainable (A)**: The goal should be attainable given the current situation such as workload and family responsibilities. The person should feel confident towards accomplishing the goal.
- **Realistic (R)**: Set goals that are not too easy, but at the same time are not too hard. The goal should appear like a challenge but still be realistic. In other words, it should be more challenging than the current average, but not too far off.
- **Time-Bound (T)**: Set an upper-bound time by which you want to achieve the goal. It is the higher level measurable component that is not set regularly but instead is an overall time frame.

An example of a well-defined SMART goal is, *I will walk 5000 steps three days a week on Monday, Wednesday, and Friday for 2 weeks*, where walk is a specific activity; 5000 steps and 3 days are measurable quantities; 2 weeks is the total time frame. As concerns attainability and realism, they are not immediately available from this goal statement and will depend on the person's circumstances. On the other hand, *I will start walking more* is a poorly defined, vague, and unquantified goal and is not likely to lead to success.

4 Data Collection and Analysis

Dataset 1: We recruited 28 patients between the ages of 21 to 65 years who were interested in increasing their physical activity at our university's internal medicine clinic. A health coach, trained in SMART goal setting, conversed with the patients to set goals every week for four weeks via

	1	2	3	4	5	6	7	8
Dataset 1								
C	19.6	13.8	12.3	11.8	-NA-			
P	15.3	12.0	10.7	11.0	-NA-			
T	34.9	25.8	23.0	22.8	-NA-			
Dataset 2								
C	14.9	14.2	11.1	11.3	8.7	9.5	9.2	9.2
P	11.0	10.9	7.8	7.1	5.6	6.9	6.6	6.7
T	25.9	25.1	18.9	18.4	14.3	16.4	15.8	15.9

Table 1: Average number of messages per week
T: total, C: Coach, P: Patient

text messages. Each week wasn't necessarily 7 days as sometimes patients took longer to set a goal, which made some weeks shorter like 5-6 days and some longer like 8-9 days. Patients used their smartphones and texting service to communicate with the coach. The coach used a web application named Mytapp, developed by Dr. Ben Gerber, to send texts to the patients. The application has been used to conduct other text-based health monitoring studies (Stolley et al., 2015; Kitsiou et al., 2017). Mytapp is a two-way text messaging application that was designed to help promote healthy behaviors and manage chronic diseases. The main benefit of using it over a normal texting service is the privacy of data. All data is encrypted and exchanged using transport layer security. The messages were saved in a secured database and the application stored minimum information about the patients.

The patients were also given Fitbits to monitor their progress. The coach monitored patients' progress using the Mytapp application, as it can fetch the most up-to-date activity data from a patient's Fitbit account and show it at one place along with text messages. This reduces the workload for the coach as at any point in time during the study the coach had at least 3 patients and would have had to login into their respective accounts to access the Fitbit data without the application. The coach needed all this information to help patients set realistic goals based on previous weeks' performance.

The HC conversations involved setting a specific, measurable and realistic goal, and solving any barriers to goal attainment. The coach sent reminders based on patients' preferences and provided motivational feedback on their progress. Out of 28 patients, only one did not finish the study due to health problems. Therefore, we only considered 27 patients' data for analysis and building models. Dataset 1 comprises 2853 messages, where 54% of messages were sent by the coach and 46% by the patients. This tells us that both the coach and the patients were equally involved. An excerpt was shown earlier in Figure 1.

Lessons from dataset 1 collection: During the initial face-to-face recruitment process at the university clinic, patients were given information about the study and the concept of SMART goal setting was explained to them. To help them understand it clearly, the goal for the first week was sometimes discussed during that initial interaction. Hence, we found that portions of the initial goal setting conversation may have been missing from the text messages, including: the patient's goal for the first week, discussion that led to that goal, and any time preferences for the text messages. Therefore, during dataset 2 collection, we asked the recruiters to take notes about what was discussed face-to-face, and asked health coaches to reiterate the first goal in text messages even if it was already known. In cases where patients didn't have information about their current activity level, a goal of one mile (2000 steps) a day was suggested. The recruiters also helped patients with setting up Fitbit trackers, downloading the Fitbit app, and linking the two together during the initial recruitment process (same as dataset 1). However, based on dataset 1 collection, recruiters had a better understanding of the issues that might arise with Fitbit and also met the patients again during the study (if possible) to fix the issues. Lastly, during the dataset 2 collection, we also collected audio-recorded feedback at the end of the study if the patients came back to the clinic, else feedback was taken over a phone call and notes were recorded.

Dataset 2: We recruited three different individuals trained in SMART goal setting to be health coaches. We also recruited 30 different patients and conducted the study for eight weeks instead of four to analyze changes in messaging behavior over a longer period. The same Mytapp application was used to text the patients and Fitbits were given to the patients. Out of 30 patients, one patient withdrew after 5 weeks, one lost their Fitbit after 2 weeks, and one set goals for only 2 weeks and then almost stopped responding. Since the latter two were in the study for fewer than 4 weeks, we only consider the data from 28 patients. We also removed all the messages discussing an appointment time for the exit interview, which comprises more than 600 messages. This resulted in 4134 messages among which 58% were sent by coaches and 42% by the patients. Dataset 1 only included about 30

Stage	Phase	Description	Phase Boundary Examples
Goal Setting	Identification	during the beginning of the week when the coach asks the patients about their goal or when the patients inform their goal to the coach	Coach: Now what goal could you make that would allow you to do more walking?
	Refining	when the coach asks (or the patient informs) the specifics of the goal such as time, location, frequency to make the goal more effective	Coach: what time do you plan to do so I can set up a reminder?
	Anticipate Barrier	when the coach asks the patients (or the patients specify) their confidence in achieving the goal (range 1-10) or if they see any upcoming barriers	Coach: Do you think the weather will make it hard for you to take 50 min walks everyday this week?
	Solve Barrier	when the coach tries to help patients overcome a barrier or increase their attainability score to 10 without modifying the quantity	Coach: what do you think will make it easy to accomplish/achieve your goal?
	Negotiation	when the patient chooses a goal that the coach thinks might be too much/less or vice-versa	Coach: another 8 . What if you were to try for 8000 steps again this week would the answer be a 10?
Goal Implementation	Refining	same as the previous stage; here it usually follows solve barrier or goal negotiation phase to make the goal more specific	Coach: Have you decided when you would like to get your walk in?
	Anticipate Barrier	similar to the previous stage, but here it indicates the barrier that has been encountered	Patient: Good morning [NAME]. I probably won't be able to make my goal this week. I'm at a professional development all day today and there are no stairs in this building
	Solve Barrier	same as the previous stage	Coach: Do you want to try to make your goal over the weekend?
	Negotiation	when the patient is unable to accomplish the goal or wants to do more, the coach or the patient asks to modify the goal	Patient: Please change my safety goal to three days per week.
	Follow up	when the coach asks the patient (or patients themselves inform) about their progress and if they can accomplish the goal	Coach: Good afternoon! How is your goal for this week going so far?

Table 2: Stages and phases schema description with examples

messages in total from 2 patients regarding appointment. So we didn't eliminate them.

Table 1 shows the average number of messages exchanged weekly, where a week corresponds to the patient's goal. There is a decrease in the number of messages over the weeks. This is because during the first week the coach sometimes redefines what a SMART goal is and also explicitly asks the patients to specify details such as which day, what time, and how much. However, as the study progresses, the answers to some of these questions such as time and days are implicitly understood to be the same as in the previous weeks if not stated otherwise and only the amount of activity is modified. Dataset 2 was collected two years after dataset 1 and hence the schemas and models were built using dataset 1 exclusively without any bias from dataset 2.

5 Annotation of the Coaching Dialogues

In this section, we will look at the two types of annotations: SMART goal annotations and stages-phases annotations. Since no work exists that has used SMART criteria to set physical activity goals via SMS, we designed the schemas that were inspired by the literature on goal setting (Bodenheimer et al., 2007; Bovend'Eerdt et al., 2009). We used the General Architecture for Text Engineering (GATE) tool for annotations (Cunningham, 2002).

Stages and Phases Annotation Schema: 15 patient-coach conversations from dataset 1 were used to design stages-phases schema. This annotation aims to understand how the conversation unfolds in HC dialogues. Stages and phases respectively help to capture the coaching tasks and sub-tasks being performed throughout the communication dialogue. The annotation schema along with descriptions is shown in Table 2. The higher tier is composed of stages; Goal Setting (GS) and Goal Implementation (GI). Stages are composed of phases. The GS stage consists of identification, refining, negotiation, anticipate barrier, and solve barrier. The GI stage consists of the same phases plus an additional follow up phase and minus the identification phase. We annotated the first message that indicated a change in a phase and all the messages after that are assumed to belong to that phase until there is a change in phase. Each message belongs to only one stage-phase. A snippet of

Stage: Goal Setting

Phase: Goal Identification
Coach: What would you like to set as your SMART goal this week?
Patient: Smart goal 12k steps a day?

Phase: Goal Negotiation
Coach: Ok, just something to think about... You got 12K steps 3 out of 7 days in the last week. That was Saturday, Sunday and Monday. How many days out the week do you want to do 12K step? Everyday?
Patient: Let's do 15K
Coach: That's more
Patient: 12k TU ,W, TH
Coach: Are you sure? If you think 12K everyday is realistic for you , go for it!
Patient: It's a challenge I'll try
Coach: Let's keep it at Tue, Wed. and Thurs then.
Patient: Ok

Phase: Solve Barrier
Coach: what do you think will make it easy to accomplish/achieve your goal?
Patient: Use stairs more and less elevator

Phase: Anticipate Barrier
Coach: On a scale of 1-10 with 10 being very sure and 1 not at all sure. How sure are you that you will accomplish your goal?
Patient: 8
Coach: What do you think will make it difficult?
Patient: Not being able to walk during my lunch hours because it's busy at work. So Time.

Phase: Goal Negotiation
Coach: I see maybe you should pick weekend days. That's when you have been most active according to fitbit
Coach: Last Sat and Sunday you got well over 12K steps
Coach: or maybe cut down on the amount of steps on those days. How can you change your goal to make that a 10 on the scale?
Patient: Ok. 12k on weekends
Coach: Sounds great good luck!!

Stage: Goal Implementation

Phase: Follow up
Coach: Good morning! How is your goal for this week going so far?
Patient: Good morning. It's going great

Figure 2: Example showing usage of stages and phases annotation schema

Stage	Phase	Message Count	Boundary Count
Goal Setting	Identification	408	109
	Refining	344	85
	Anticipate Barrier	363	82
	Solve Barrier	158	52
	Negotiation	92	19
Goal Implementation	Refining	16	4
	Anticipate Barrier	8	4
	Solve Barrier	25	7
	Negotiation	23	6
	Follow up	1348	120

Table 3: Stage-phase tags. Number of: messages in given stage-phase ('Message count'); dialogue transitions into given stage-phase ('Boundary count').

	gs_I	gs_R	gs_N	gs_AB	gs_SB	gi_R	gi_N	gi_AB	gi_SB	gi_F	Stop
Start	1	0	0	0	0	0	0	0	0	0	0
gs_I	0	0.63	0.04	0.17	0.08	0	0	0	0	0.08	0
gs_R	0.01	0	0.06	0.38	0.2	0	0	0	0	0.35	0
gs_N	0	0.05	0	0.21	0.16	0	0	0	0	0.58	0
gs_AB	0	0.05	0.1	0	0.28	0	0	0	0	0.57	0
gs_SB	0.02	0.21	0.04	0.54	0	0	0	0	0	0.19	0
gi_R	0	0	0	0	0	0	0.25	0	0	0.75	0
gi_N	0	0	0	0	0	0.33	0	0	0	0.67	0
gi_AB	0	0	0	0	0	0	0.5	0	0.25	0.25	0
gi_SB	0	0	0	0	0	0.14	0	0	0	0.71	0.14
gi_F	0	0	0	0	0	0.01	0.02	0.03	0.05	0	0.88

Figure 3: Transition probabilities from one stage-phase to another [gs: goal setting, gi: goal implementation, I: identification, R: refining, N: negotiation, AB: anticipate barrier, SB: solve barrier, F: follow-up]

an annotated conversation is shown in Figure 2.

Two annotators annotated four previously unseen patients' data for stages and phases (447 messages). Inter Annotator Agreement (IAA) was measured using Cohen's kappa coefficient (κ) (Cohen, 1960); we obtained an excellent κ=0.93. This may be partially due to the stages-phases being bound to occur in a particular sequence: our HC conversations follow a particular structure, which involves phases such as identification, refining, and negotiation. Therefore, we analyzed the HC conversations as concerns likely transitions, and their frequencies.

First, Table 3 shows the counts for stage-phase

Tag	Feature	Description	Slot Example	Intent Example
Specificity	activity	the activity that will be done by the patient	Patient: Ok. I'll walking the **stairs** in the mornings from 8 to 10 Monday - Friday	
	time	the time of the day when the patient will be doing the activity	Patient: Ok. I'll walking the stairs **in the mornings from 8 to 10** Monday - Friday	Coach: like how many days next week and **at what time of day?**
	location	the location where the patient will be doing the activity	Patient: I can also plan to walk the stairs **at home**. After work	Coach: **Could you maybe get your steps done in the house?**
Measurability	quantity (amount/ distance/ duration)	quantifies the activity in some way to show what patients are planning to accomplish. It can be number of steps or stairs, distance or duration	Patient: Yes, I'm going for **6000 step** *(amount)* Patient: I will walk **3 blocks** *(distance)* Patient: I do **40 min** of walk *(duration)*	Coach: **How many would you like to try for?**
	days (name/ number)	the number of days or the name of the days the patient will be working on the chosen activity	Patient: Ok. I'll walking the stairs in the mornings from 8 to 10 **Monday - Friday** *(days-name)* Patient: I will try **3 days**. *(days-number)*	Coach: **Will you walk 4 block on the same days Mon, Wed and Fri?***(days-name)* Coach: like **how many days next week** and at what time of day? *(days-number)*
	repetition	the number of times the activity will be done in the same day	Patient: I will attempt to spend 15 mins **3times a day** walking up and down two flights of stairs. 8am	
Attainability	score	specifies how confident a patient is about accomplishing the goal on a scale of 1-10	Patient: **8**	Coach: **On a scale of 1-10 with 10 being very sure and 1 not at all sure. How sure are you that you will accomplish your goal?**
Realism		helps to indicate statements that judge the realism of a goal for the patient. It is usually based on their previous performance		Coach: **Sounds like a very doable goal you are averaging over 9k steps during the weekdays, now**

Table 4: SMART annotation schema description with examples

Coach: What goal could you make that would allow you to do more walking?
Patient: Maybe walk (*S_activity*) more in the evening after work (*S_time*).
Coach: Ok sounds good. $\underbrace{How\ many\ days\ after\ work(S\_time)\ would\ you\ like\ to\ walk\ (S\_activity)?}_{\text{M\_days\_number\_intent}}$
Coach: $\underbrace{And\ which\ days\ would\ be\ best?}_{\text{M\_days\_name\_intent}}$
Patient: 2 days (*M_days_number*). Thursday (*M_days_name*), maybe Tuesday (*M_days_name_update*)
Coach: $\underbrace{Think\ about\ how\ much\ walking\ (S\_activity)\ you\ like\ to\ do\ for\ example\ 2\ block\ (M\_quantity\_distance\_other)}_{\text{M\_quantity\_intent}}$
Patient: At least around the block (*M_quantity_distance*) to start.
Coach: $\underbrace{On\ a\ scale\ of\ 1-10\ with\ 10\ being\ very\ sure.\ How\ sure\ are\ you\ that\ you\ will\ accomplish\ your\ goal?}_{\text{A\_intent}}$
Patient: 5 (*A_score*)

Figure 4: Example showing usage of SMART goal annotation schema

annotations in dataset 1 minus the first two to three introduction messages about the study in each conversation as they were the same. Other than *follow up*, all other phases in GI stage rarely occur.

Focusing on transitions now, a priori, 121 different transitions are possible in a given week, as we have 10 unique stage-phase categories plus the beginning and end of the week (start, stop). However, only 39 unique transitions occur in our dataset, given that a week always starts with the goal setting stage, which in turn starts with the goal identification phase. On further analysis, we found that only 13 of those 39 transitions have a probability above 0.3, as shown in Figure 3.

SMART Tag Annotation Schema: Similar to stage-phase annotations, 15 patient-coach conversations were used to design the SMART goal annotation schema. The schema is described in Table 4

Level	S	M	A	R
Message	0.967	0.965	**0.907**	**0.694**
Word	**0.878**	**0.895**	0.515	0.549

Table 5: Kappa on SMART goal annotation schema

Tag	Feature	Slot Value	Intent
Specificity	Activity	671	0
	Time	131	31
	Location	41	1
Measurability	Quantity	627	30
	Days	303	63
	Repetition	69	0
Attainability		70	261
Realism		N/A	70

Table 6: Counts for SMART tags in the dataset 1

along with examples. We didn't annotate for Timeliness as a new goal was set every week, and hence by default, its value is one week. Each annotation can either be categorized as a slot value or an intention. A slot value is a word or group of words that capture a particular piece of information, for example, 'walk' is a slot value for *specific activity*; the intention is an utterance that tries to gain information about a slot. Each SMART annotation category can have other optional tags such as *previous* to annotate an attribute related to the previous week's goal, *accomplished* or *remaining* to annotate the progress of the patient, *update* to add another slot value to an existing one, and *other* for anything which doesn't belong to the previous or current week. Figure 4 shows the use of the SMART annotation schema.

Two annotators annotated four previously unseen patients' data for SMART goal attributes. Results for IAA measured using kappa (κ) is shown in Table 5. We measured κ on two levels: message and word. At the message level, we consider an agreement if both the annotators labeled at least one word in the message with the given tag (not necessarily the same word). At the word-level, we consider it an agreement if both annotators labeled the same word with the given tag.

In total, 447 messages were annotated for IAA. There were 128 messages with Specificity (S) tag, 120 with Measurability (M) tag, 45 with Attainability (A) tag and 13 with Realism (R) tag. We achieved $\kappa \approx 0.9$ for {S, M} and $\kappa \approx 0.5$ for {A, R}. This is mostly because {S, M} tags have a higher number of occurrences in the data as compared to {A, R} which are hard to distinguish from each other and have very few occurrences. It should

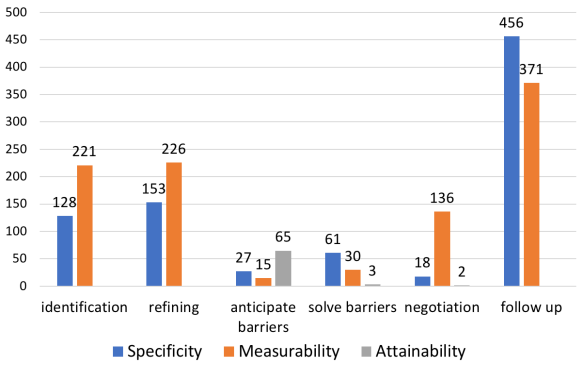

Figure 5: SMART tag counts per phase

also be noted that for {S, M} word-level annotation is more important whereas for {A, R} message level annotation makes more sense. Table 6 shows the counts for SMART categories in dataset 1. One can notice that the percentage of {R} is fairly small as compared to the {S, M, A} tags. It is not surprising as the coach only questions the realism of the goal if he thinks it is either too difficult/easy based on the patient's past performances.

6 Development on Dataset 1

Dataset 1 has been our foundation to develop the computational models we are interested in, namely SMART tag and phase prediction. Before building these models, we wanted to see if SMART tags and phases exhibit any sort of relationship that can be leveraged as features. We plotted the number of SMART tags in each phase and obtained the graph shown in Figure 5. SMART tags are unevenly distributed across phases, with identification, refining and follow up containing the majority of SMART tags. Therefore, we experimented with SMART tags as a feature in the phase prediction model and vice versa, and found that SMART tags helped to predict phases better, than phases help predict SMART tags (Gupta et al., 2019).

We achieved an average (macro) F1 score of 0.80 on SMART tag prediction using Structured Perceptron with feature combination of the current and surrounding words, pre-trained Google word embeddings[3], and SpaCy[4] named entity recognizer output. Similarly, we achieved an average (macro) F1 score of 0.71 on phase prediction using Conditional Random Fields with feature combination of unigrams, distance of the message from the top in a given week, and human-annotated SMART tag

[3] https://code.google.com/archive/p/word2vec/

[4] https://spacy.io/

(1) Coach: Okay, so your goal this week is to reach 17,500 steps (*M_quantity_amount*) one day (*M_days_number*) this week (Monday through Sunday) (*M_days_name*), correct? (**Human and Automated**)

(2) Coach: Your goal is to reach 10,000 steps (*M_quantity_amount*) any day (*M_days_number*) this week by Friday (*M_days_name*)! You said that you give your confidence a 9 (*A_score*) on a 10 point scale. You can do this! (**Human**)

(2) Coach: Your goal is to reach 10,000 steps (*M_quantity_amount*) any day this week by Friday (*M_days_name*)! You said that you give your confidence a 9 (*A_score*) on a 10 point scale. You can do this! (**Automated**)

Figure 6: Automated annotation output (dataset 2).

counts. Importantly, an almost similar performance (F1 score = 0.69) was achieved using automatically predicted SMART tags.

Unfortunately, use of deep learning is not suitable due to our very small dataset; only 2853 messages in total. One can also notice rare occurrences of classes such as anticipate barriers, solve barriers, and negotiation in Figure 5.

7 Applications of Models Developed on Dataset 1

So far the models developed in Section 6 have been used for two applications: annotating dataset 2 for both SMART tags and phases, and goal extraction.

Goal extraction on dataset 1: Goal extraction can help health coaches to recall a goal discussed during the conversation and save their time. Since SMART tags helped predict phases better, we built a pipeline where SMART tags were predicted first, then they were used as one of the features in phase prediction. After SMART tag and phase prediction, we extracted the SMART tags as long as they were not from the follow-up phase to avoid extracting accomplished and remaining *measurable quantity*. 65% of the goals we extracted correctly identified at least 8 out of 10 SMART attributes of the gold standard goal. Detailed results for goal extraction and the two models are available in Gupta et al. (2020). We are currently evaluating our goal extraction model on dataset 2 with the help of health coaches and automatic evaluation.

Dataset 2 annotation: We used the same pipeline for annotating dataset 2, except we changed Google word embeddings to pre-trained ELMo word representations for SMART tag prediction (Peters et al., 2018). To measure performance, we manually annotated three randomly chosen patients' data, one from each coach. We achieved an F1 score of 0.81 (macro) and 0.98 (weighted) on SMART tag annotations and 0.37 (macro) and 0.61 (weighted) on phase annotations. The results for SMART tag prediction on dataset 2 is equal to what we achieved on dataset 1. This means that SMART tag annotations are transferable even if the dialogues are between different coaches and patients. A sample output for SMART tags is shown in Figure 6. Our model correctly annotated (1), but missed *M_days_number* in (2). More specifically, for the three patients that we automatically annotated, only 113 words (2%) were incorrectly labelled or had a missing label; 390 words (6%) were correctly labelled with a SMART tag; and 5959 words (92%) were correctly labelled with 'none' tag.

Because performance on automatic phase annotation was not as high as we had hoped, we will adopt a semi-automatic approach, with a round of manual edits following automatic annotation of phases. We see semi-automatic annotation as crucial, especially given that state-of-the-art deep learning models require large labeled training data. Semi-automatic annotation can still save thousands of hours of manual labor.

8 Conclusions and Future Work

We envision a virtual assistant health coach that can help people to increase their physical activity by motivating them to set SMART goals. To this end, we collected a health coaching dialogue dataset and developed two annotation schemas, one that captures the slot-values of a SMART goal and the other that captures the higher-level conversation flow of the health coaching dialogues. We briefly discussed the models built using the two annotations and their application for automatic goal extraction. We also collected a second round of dataset and showed that it can be reliably annotated using the models built on the first dataset. Our immediate next steps are to perform extrinsic evaluation of the goal extraction pipeline with the help of our health coaches and integrate it into the Mytapp application used by the health coaches for round three of the data collection.

9 Acknowledgements

This work is supported by the National Science Foundation through awards IIS 1650900 and 1838770.

References

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: A corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Timothy W Bickmore, Daniel Schulman, and Candace L Sidner. 2011. A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. *Journal of Biomedical Informatics*, 44(2):183–197.

Thomas Bodenheimer, Connie Davis, and Halsted Holman. 2007. Helping patients adopt healthier behaviors. *Clinical Diabetes*, 25(2):66–70.

Frank W Booth, Christian K Roberts, John P Thyfault, Gregory N Ruegsegger, and Ryan G Toedebusch. 2017. Role of inactivity in chronic diseases: evolutionary insight and pathophysiological mechanisms. *Physiological Reviews*, 97(4):1351–1402.

Fiemke Both, Pim Cuijpers, Mark Hoogendoorn, Michel CA Klein, A Fred, J Filipe, and H Gamboa. 2010. Towards fully automated psychotherapy for adults: BAS-behavioral activation scheduling via web and mobile phone.

Thamar JH Bovend'Eerdt, Rachel E Botell, and Derick T Wade. 2009. Writing smart rehabilitation goals and achieving goal attainment scaling: a practical guide. *Clinical Rehabilitation*, 23(4):352–361.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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.

Ester Cerin, Evie Leslie, Takemi Sugiyama, and Neville Owen. 2010. Perceived barriers to leisure-time physical activity in adults: an ecological perspective. *Journal of Physical Activity and Health*, 7(4):451–459.

Ananlada Chotimongkol and Alexander I Rudnicky. 2008. Acquiring domain-specific dialog information from task-oriented human-human interaction through an unsupervised learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 955–964. Association for Computational Linguistics.

Clara K Chow, Julie Redfern, Graham S Hillis, Jay Thakkar, Karla Santo, Maree L Hackett, Stephen Jan, Nicholas Graves, Laura de Keizer, Tony Barry, et al. 2015. Effect of lifestyle-focused text messaging on risk factor modification in patients with coronary heart disease: a randomized clinical trial. *JAMA*, 314(12):1255–1263.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Hamish Cunningham. 2002. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254.

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

Nishitha Guntakandla and Rodney Nielsen. 2018. Annotating reflections for health behavior change therapy. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Bing Liu, Ben Gerber, and Lisa Sharp. 2019. Modeling health coaching dialogues for behavioral goal extraction. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1188–1190. IEEE.

Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Bing Liu, Ben Gerber, and Lisa Sharp. 2020. Goal summarization for human-human health coaching dialogues. In *Florida Artificial Intelligence Research Society Conference*.

Margaret Handley, Kate MacGregor, Dean Schillinger, Claire Sharifi, Sharon Wong, and Thomas Bodenheimer. 2006. Using action plans to help primary care patients adopt healthy behaviors: a descriptive study. *The Journal of the American Board of Family Medicine*, 19(3):224–231.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272.

Spyros Kitsiou, Manu Thomas, G Elisabeta Marai, Nicos Maglaveras, George Kondos, Ross Arena, and Ben Gerber. 2017. Development of an innovative mhealth platform for remote physical activity monitoring and health coaching of cardiac rehabilitation patients. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 133–136. IEEE.

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.

Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. 2016. The motivational interviewing treatment integrity code (miti 4): rationale, preliminary reliability and validity. *Journal of Substance Abuse Treatment*, 65:36–42.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 42–51.

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 2018. Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Julia Poncela-Casasnovas, Bonnie Spring, Daniel McClary, Arlen C Moller, Rufaro Mukogo, Christine A Pellegrini, Michael J Coons, Miriam Davidson, Satyam Mukherjee, and Luis A Nunes Amaral. 2015. Social embeddedness in an online weight management programme is linked to greater weight loss. *Journal of The Royal Society Interface*, 12(104):20140686.

Melinda R Stolley, Lisa K Sharp, Giamila Fantuzzi, Claudia Arroyo, Patricia Sheean, Linda Schiffer, Richard Campbell, and Ben Gerber. 2015. Study design and protocol for moving forward: a weight loss intervention trial for african-american breast cancer survivors. *BMC Cancer*, 15(1):1018.

Alice Watson, Timothy Bickmore, Abby Cange, Ambar Kulshreshtha, and Joseph Kvedar. 2012. An internet-based virtual coach to promote physical activity adherence in overweight adults: randomized controlled trial. *Journal of Medical Internet Research*, 14(1).

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413.