Towards a Virtual Assistant Health Coach: Corpus Collection and Annotations

Itika Gupta\textsuperscript{1} and Barbara Di Eugenio\textsuperscript{1} and Brian Ziebart\textsuperscript{1} and Bing Liu\textsuperscript{1} and Ben Gerber\textsuperscript{2} and Lisa Sharp\textsuperscript{2} and Rafe Davis\textsuperscript{2} and Aiswarya Baiju\textsuperscript{1}
\textsuperscript{1} Department of Computer Science, \textsuperscript{2} Institute for Health Research and Policy
University of Illinois at Chicago

1 Introduction

Many health problems faced by individuals can be mitigated with changes in health behavior. However, successfully implementing healthy behaviors in ones daily life requires significant motivation that most people, individually, find difficult to initiate and maintain. Health coaching has been identified as a successful method for motivating and maintaining health behavior changes by having a peer or professional convey relevant medical information, help to set realizable, yet challenging goals tied to health behavior change, and provide encouragement in adhering to those goals (Kivelä et al., 2014). But unfortunately, personal health coaching is time-intensive, uneconomical for the low-income patients, and has limited accessibility.

Therefore, we aim to create a dialogue-based virtual assistant health coach that will converse with the patient via text messages and will help them increase physical activity by setting Specific, Measurable, Achievable, Realistic and Time-bound (SMART) goals (Doran, 1981). Even though an influential strand of work on conversational agents has been conducted by Bickmore and his group, their systems rely on predefined set of utterances from the patients (Bickmore et al., 2015). We plan to build an autonomous system that learns from observed communications between human health coach and participants to increasingly automate the generation of dialogue (Rizzolatti and Craighero, 2004). The ability to recognize proposed health goals in dialogues is a key initial capability needed for this type of learning from demonstration.

In this paper, we talk about our data collection process, two annotation schemas, agreement results and future work on extracting goals from patient-coach dialogue.

2 Data Collection

We recruited 28 patients between the age of 21 to 65 years who were interested in increasing their physical activity. A health coach, trained in SMART goal setting, conversed with these patients to set goals on weekly basis for a month via SMS. The patients were given Fitbit Alta to monitor their progress. The coach also monitored patients’ progress using Fitbit application. The conversation involved setting a specific, measurable and realistic goal, and establishing any barriers to goal attainment. The coach also sent reminders based on patient’s preference and provided motivational feedback on their progress. Only one patient didn’t complete the study. We have a corpus of 2857 messages. Among these messages, approximately 53% were sent by the coach and 47% by the patients. This tells us that both the coach and the patients were equally involved in setting or negotiating the goal.

3 Data Annotation

In this section we will look into two types of annotation schemas designed by us to help automate the extraction of health goal set by the user.

3.1 SMART Goal Annotation

3.1.1 Annotation Schema

15 patient-coach conversations were used to design the SMART goal annotation schema shown in Figure 2 (Bovend’Eerdt et al., 2009). We didn’t annotate Time-liness as a new goal was set every week, and hence by default its value is one week. Each of the annotations can either be categorized as a slot value or an intention. A slot value is a word or group of words that captures a particular piece of information such as ‘walk’ is a slot value for specific activity. Where as intention is a utterance that tries to gain information about a slot. Each of the SMART annotation category can have other optional tags such as previous to annotate an attribute related to 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 previous or current week. Figure 1 shows the use of SMART Annotation schema.

3.1.2 Annotation Results

Two annotators annotated four previously unseen patients’ data for SMART goals. Inter annotator agreement (IAA) was measured using kappa statistics on individual SMART categories and the results are shown in Table 1.

We measured kappa on two levels: message and word. In message level, we consider an agreement if both the annotators labeled at least one word in the
Figure 1: SMART goal annotations for a conversation between the health coach and the patient.

Figure 2: SMART goal annotation schema

<table>
<thead>
<tr>
<th>Level</th>
<th>S</th>
<th>M</th>
<th>A</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Message</td>
<td>0.967</td>
<td>0.965</td>
<td>0.907</td>
<td>0.694</td>
</tr>
<tr>
<td>Word</td>
<td>0.878</td>
<td>0.895</td>
<td>0.515</td>
<td>0.549</td>
</tr>
<tr>
<td>Table 1: Results for Inter-Annotator Agreement.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

message with the given tag (not necessarily the same word). In word level, we consider an agreement if both the annotators labeled the same word with the given tag.

In total, 447 messages were annotated for IAA. There were 128 messages with Specificity tag, 120 with Measurability tag, 45 with Attainability tag and 13 with Realism tag. We observe approximately 90% reliability for Specificity and Measurability and only 50% reliability for Attainability and Realism. This is because \{S, M\} tags are easy to annotate and have high 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 also be noted that for \{S, M\} word level annotation is more important while for \{A, R\} message level annotation makes more sense.

Table 2 shows the statistics for SMART categories over the entire dataset of 2857 messages. One can observe that the percentage of \{R\} in the entire dataset is fairly small when compared to the \{S, M, A\} tags. It is not surprising as the coach only questions the realism of the goal if he thinks the goal is either too difficult or too easy based on the patient’s past performances.

3.2 Stages and Phases Annotation

3.2.1 Annotation Schema

The aim of this annotation is to understand how the conversation unfolds in a health coaching dialogue. Stages and phases help to capture the coaching tasks and sub tasks being performed throughout the communication dialogue respectively. The higher tier is composed of stages; goal setting and goal action. Stages are composed of phases. Goal setting stage consists of goal identification, goal refining, goal negotiation, anticipate barrier and solve barrier. Goal action stage consists of the same phases plus an additional follow up phase. A snippet of such conversation is shown in Figure 3. Each message can be annotated as goal setting, goal action, or none. Phases are optional.

3.2.2 Annotation Results

Similar to SMART annotation, two annotators annotated four previously unseen patients for stages and phases. The confusion matrix showing the annotation counts is shown in Table 3, where ‘gs’ stands for goal setting, ‘ga’ stands for goal action, ‘I’ stands for goal identification, ‘R’ for goal refining, ‘A’ for anticipate barrier, ‘S’ for solve barrier, ‘N’ for goal negotiation, and ‘F’ for follow up. The kappa statistics for all the categories together came out to be 0.93.

In total, 398 messages were annotated for IAA. Out of them, 115 messages were annotated as goal setting whereas 59 messages were annotated as goal action by both the annotators. The high value of kappa can be attributed to the fact that while following SMART criteria for goal setting, the stages/phases are bound to occur in a particular order and therefore, easy to annotate.

Table 4 shows the summary of stages and phases annotations over the entire dataset of 2857 messages. Many categories have zero occurrence in the data used for IAA as they are not very frequent in the entire dataset.
Coach: What goal could you make that would allow you to do more walking? (Goal Setting - Goal Identification)
Patient: I could walk on my lunch break for 20 mins Wed and Thurs (Goal Setting - Goal Identification)
Coach: Good answer
Coach: What do you think will make it easy to accomplish/achieve your goal? (Goal Setting - Anticipate Barrier)
Patient: I just have it set in my mind... And once I set a goal I like to see it through (Goal Setting - Anticipate Barrier)
Coach: I can also send some reminders. What time do you take lunch? (Goal Setting - Goal Refining)
Patient: Ok that’s sounds good... Normally around 1030-11 (Goal Setting - Goal Refining)
Coach: Sounds like a plan. Good luck! (Goal Action)
Patient: Thanks
Coach: Don’t forget to walk on lunch break
Patient: Ok perfect timing...I’m going on break now
Coach: Good morning! How is your goal for this week going so far? (Goal Action - Follow up)
Patient: Wed I only walked 10 mins of my break due to the rain. Thur I did the full 20min (Goal Action - Follow up)
Coach: You can get another 10 min in today or Monday so you can still accomplish your goal. (Goal Action - Solve Barrier)
Coach: Good morning! Were you able to get that 10 minute walk in last week? (Goal Action - Follow up)
Patient: Yes everything worked out great (Goal Action - Follow up)

Figure 3: Stages and Phases Annotations for a Conversation between the Health Coach and the Patient.

<table>
<thead>
<tr>
<th>gs_I</th>
<th>gs_R</th>
<th>gs_A</th>
<th>gs_S</th>
<th>gs_N</th>
<th>ga_J</th>
<th>ga_R</th>
<th>ga_A</th>
<th>ga_S</th>
<th>ga_N</th>
<th>ga_F</th>
<th>ga</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>210</td>
</tr>
</tbody>
</table>

Table 3: Confusion Matrix for Stages and Phases Annotations.

<table>
<thead>
<tr>
<th>gs_I</th>
<th>gs_R</th>
<th>gs_A</th>
<th>gs_S</th>
<th>gs_N</th>
<th>ga_J</th>
<th>ga_R</th>
<th>ga_A</th>
<th>ga_S</th>
<th>ga_N</th>
<th>ga_F</th>
<th>ga</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>252</td>
<td>245</td>
<td>291</td>
<td>25</td>
<td>36</td>
<td>1</td>
<td>15</td>
<td>7</td>
<td>8</td>
<td>12</td>
<td>217</td>
<td>102</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Messages in each Stage and Phase Category

4 Discussion and Future Work

In this paper we discussed our data collection process, SMART goal annotation schema, and stages and phases annotation schema. We plan to use these annotations to train a classifier that can extract and summarize the goal set by the patient. We are currently working on a pipeline that will first detect all the possible SMART goal features in the conversation, then detect the stages/phases in the conversation and then finally combine these two to extract the correct SMART goal. For example, if a goal was set during the goal identification phase, but then due to some barrier the goal changed in the goal negotiation phase, the summarization model should summarize the updated goal.

References


