Goal Summarization for Human-Human Health Coaching Dialogues

Itika Gupta,[1] Barbara Di Eugenio,[1] Brian Ziebart,[1] Bing Liu,[1] Ben Gerber,[2] Lisa Sharp[2]
{igupta5, bdieugen, bziebart, liub, bgerber, sharpl}@uic.edu
[1]Department of Computer Science
[2]Institute for Health Research and Policy
University of Illinois at Chicago

Abstract

Lack of physical activity has been linked to several chronic diseases. Health coaching is successful to help patients engage in healthier behaviors, but is resource intensive. Our goal is to develop a virtual health coach. In this paper, we discuss one component of our work, automatically summarizing goals set by patients during health coaching conversations that we collected and annotated. In turn, our goal summarization pipeline consists of a slot-value prediction model followed by a model that captures the higher-level conversation flow of the dialogues. We report a detailed evaluation that shows measures used for summarization such as BLEU and ROUGE, do not work well for our task.

Introduction

Physical inactivity is a primary reason for many chronic diseases such as type 2 diabetes, cardiovascular disease, and depression (Booth et al. 2017). According to the Physical Activity Guidelines for America, an individual should do at least 150 minutes per week of moderate-intensity physical activity to be considered active (Piercy et al. 2018). But unfortunately, only 22.9% of the United States (U.S.) adult population met the federal guidelines for physical activity between the years 2010 - 2015 (Blackwell and Clarke 2018). Therefore it comes as no surprise that every 6 out of 10 adults in the U.S. suffers from at least one chronic disease.

Many of these problems can be reduced by increasing the amount of physical activity. However, the problem is being able to maintain these activities regularly and the continuous motivation needed to do so. 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). Unfortunately, personal HC is time-intensive, too expensive for low-income patients, inflexible in terms of availability of the coach, and may have limited reach because of distance, especially for people from rural communities.

In the last few years, there has been a considerable interest in automated systems for health behavior change (Watson et al. 2012; Shamekhi et al. 2017). 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. Our goal is to develop a virtual assistant health coach that will help patients to set Specific, Measurable, Attainable, Realistic and Time-bound (S.M.A.R.T.) goals via text messages (Doran 1981).

We follow the traditional architecture for building a dialogue system since we have small data and an end-to-end system is not feasible. We first build the Natural Language Understanding (NLU) module, which involves understanding the user's intent (stages-phases) and slot values (SMART goal attributes). Whereas we will shortly use intent and slots as inputs to the dialogue manager, in the meantime, we use the NLU module to support human health coaches and provide them with automatically generated summaries of the patients' goals. Our contributions are as follows:

1. We developed a model for extracting behavioral goals discussed during human-human HC dialogues. We believe this is the first model developed for the HC domain.
2. We leveraged the conversation flow of the HC dialogues we collected to determine the slot values of the agreed-upon goal. Our slot prediction model is free of any fixed ontology as values are unbounded for most of our slots.
3. We show that standard metrics like BiLingual Evaluation Understudy score (BLEU) (Papineni et al. 2002) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin 2004), are not appropriate to evaluate our extraction based goal summaries.

Related Work

One of the earliest and simplest methods to interact with patients involves programmable prompting devices, which send reminders to the participants and are useful for interventions such as medication adherence and smoking cessation (Andrade et al. 2005). Then came conversational agents that interact with the users to help them with activities such as managing stress and assisting patients during hospital visits (Bickmore et al. 2015; Shamekhi et al. 2017). These systems are sometimes embodied via an animated character that uses both verbal and non-verbal cues such as hand gestures and eye gaze to build a rapport with the user.

However, most of these systems provide a predefined set of options for input. Though this enhances system's robust-

Stage: Goal Setting

(1) Coach: What goal could you make that would allow you to do more walking?
(2) Patient: Maybe walk (*S_activity*) more in the evening after work (*S_time*).
(3) Coach: Ok sounds good. How many days after work (*S_time*) would you like to walk (*S_activity*)?
(4) Coach: And which days would be best?
(5) Patient: 2 days (*M_days_number*). Thursday (*M_days_name*), maybe Tuesday (*M_days_name*)
(6) Coach: Think about how much walking (*S_activity*) you like to do for example 2 block (*M_quantity_distance*)
(7) Patient: At least around the block (*M_quantity_distance*) to start.
(8) Coach: On a scale of 1-10 with 10 being very sure. How sure are you that you will accomplish your goal?
(9) Patient: 5 (*A_score*)

Figure 1: SMART goal annotations for a conversation between the health coach and the patient. S: Specificity (Blue), M: Measurability (Red), and A: Attainability (Magenta), Phases: *identification* (1-2), *refining* (3-7), and *anticipate barriers* (8-9)

ness, portability is reduced (Bickmore, Schulman, and Sidner 2011). Some of the work that does provide unconstrained natural language input uses pattern matching for extracting key-phrases from the user's input (Song et al. 2013; Fitzpatrick, Darcy, and Vierhile 2017).

Very recently researchers have started to use computational methods for analyzing conversations from counseling and interventions. Work by Althoff, Clark, and Leskovec (2016) uses 80,000 counseling conversations from a crisis text line to understand which linguistic features lead to a successful or unsuccessful counseling session. On the other hand, Pérez-Rosas et al. (2017) collected a dataset of Motivational Interviewing (MI) based sessions on smoking cessation, weight loss, and medication adherence and built a model for predicting the counselor's performance based on linguistics features. More recently, the authors used public sources such as YouTube videos on MI for the same purpose (Pérez-Rosas et al. 2018).

The use of conversational agents in the health care domain is still fairly new as compared to other domains such as traveling or booking a restaurant. The abundance of data in other domains allows researchers to utilize complex neural network models. However, in the health domain, it is very resource intensive to collect real-world data; this results in very small datasets, which are in general not sufficient to train neural networks. Also, due to privacy reasons, data can only rarely be shared. These limitations led us to collect our data which is specific to our task domain of HC.

Data Collection and Annotation

We recruited 28 patients and a health coach to communicate with patients via SMS and help them set weekly SMART goals for 4 weeks. Only one patient didn't finish the study. The patients were given a Fitbit to track their progress (also accessible to the coach). This resulted in a corpus of 2853 messages. An annotated excerpt is shown in Figure 1. Two annotators labeled the data for lower-level attributes i.e. SMART tags and higher-level conversation structure, i.e. stages-phases. We reached satisfactory levels of inter-annotator agreement (kappa) on all categories (Cohen 1960). The SMART goal schema has 10 word-level tags with examples shown in parenthesis: specific activity (*walk*), specific time (*8am*), specific location (*at work*); measurable quantity amount (*2000 steps*), measurable quantity distance (*2 miles*), measurable quantity duration (*15 min*), measurable days name (*Monday, Tuesday*), measurable days number (*2 days*), measurable repetition (*2 times a day*); and attainability score between 1-10 (*9*). The coaching stages-phases schema has two stages: Goal Setting (GS) and Goal Implementation (GI). The GS stage has 5 phases: identification, refining, negotiation, anticipate barrier, and solve barrier. The GI stage has the same phases minus identification and an additional follow-up phase. Further details can be found in (Gupta et al. 2018; 2019).

Goal Extraction Framework

Most goal-oriented dialogue systems assume that a user has a predefined goal that needs to be accomplished such as reserving a restaurant. However, that is not the case in our HC dialogues. Though patients are encouraged to set their own goals, health coaches play a major role in helping patients converge on a realistic goal based on their lifestyle and previous activity patterns; therefore our dialogues involve lots of negotiation before a goal is agreed upon. Moreover, the patients sometime change their goals on encountering an unseen barrier. This causes information about the goal to be distributed over multiple messages. Therefore, we hypothesize that understanding the current stage-phase of a message can help identify these negotiations and better predict the final goal.

There are two points in the conversations where extracting and summarizing the goal offline would help coaches; one at the end of the goal-setting stage and another at the end of the goal implementation stage. We call them the forward-looking goal and backward-looking goal respectively.

A forward-looking goal doesn't include any negotiations or changes to the goal that might occur during the goal implementation stage. E.g, for the conversation in Figure 1, the summary would be 'walk around the block after work 2 days Thursday and Tuesday'. The summary at this point can be used to send automated reminders, help coaches remember the goal, and to compare against the Fitbit readings. A backward-looking goal refers to the final goal the patient worked towards including all the negotiations during the week: e.g., during the goal implementation stage, the patient may encounter a barrier on Thursday, and change the goal to Friday. In such a case, the goal summary would be 'walk around the block after work 2 days Friday and Tues-

Feature	
U	: Unigrams
D	: Distance of the message from top in a week
SMART	: SMART attribute present or not
L	: Sentence length
T	: Normalized time difference between messages
Se	: Sender of the message
WE	: Google Word Embedding

Table 1: List of features for phase prediction

Features	CRF	SP	LR
Baseline	0.18		
U	0.626	**0.672**	0.514
U+D	0.666	0.604	0.558
U+D+SMART	0.708	0.604	0.592
U+D+SMART+L	0.702	0.622	0.592
U+D+SMART+L+T	0.704	0.622	0.600*
U+D+L+T+Se	0.664	0.590	0.566
U+D+SMART+L+T+Se	0.704	0.628	0.602
All (from Table 1)	**0.710**	0.622	**0.604**

Table 2: Phase prediction F1 scores. '*' indicates significant improvement, boldface indicates highest value.

Label	P	R	F1	Support
Baseline	0.250	0.212	0.182	532.4
Anticipate barrier	0.836	0.814	0.816	72.2
Follow up	0.908	0.922	0.912	256.4
Identification	0.816	0.858	0.828	109
Negotiation	0.482	0.360	0.368	21.2
Refining	0.660	0.732	0.678	69.6
Solve barrier	0.722	0.588	0.632	34.2
Macro average	0.738	0.712	0.708	532.4

Table 3: Phase prediction results per label

day'. Summarizing the goal at this point can give coaches a better sense of the patient's goal attainability, a direction to set future goals, and comparison against the Fitbit readings.

Phase Prediction

We first analyzed the HC conversations to see if a given stage-phase is more likely to be followed by another stage-phase in a given week. In total, 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). We found only 39 unique transitions in our dataset which was expected given that the goal implementation stage cannot be followed by the goal setting stage in a given week, and the week always starts with the goal identification phase. Moreover, out of these 39 transitions, only 13 had a probability above 0.3. Therefore, we tried both sequential and non-sequential classification algorithms for the phase prediction task. For sequential algorithms, we modeled a set of messages in one week as one sequence.

We used the 80-20 rule to divide our data into training and testing and performed 5-fold cross-validation. All the messages from one patient were either kept in training data or test data to avoid any data leakage. We used supervised classification models, specifically Conditional Random Fields (CRF), Structured Perceptron (SP), Support Vector Machines (SVM), Logistic Regression (LR) and Decision Trees (DT). We did experiment with other models as well, namely, Hidden Markov Model, Naive Bayes, and K-Nearest Neighbors, but their overall performance was worse than the five classifiers mentioned above. Therefore, we performed feature ablation experiments only with CRF, SVM, SP, DT, and LR classifiers and report the results for the top three classifiers here. The naive baseline assigns the *identification* phase to the first message in a week, and the majority tag *follow up* to all other messages that week. The list of features is shown in Table 1. We also experimented with Part-Of-Speech (POS) tags, dependency parse trees, number of content words (not stopwords) in the sentence, and the previous message's word embedding. However, none of them contributed to model performance.

Results are shown in Table 2. The first '*' in a column indicates a significant improvement ($p<0.05$) over unigrams calculated using ANOVA and post-hoc Tukey tests. The next '*' in the column indicates a significant improvement over the previous significant improvement. The highest F1 score of 0.710 was achieved using all the features with the CRF model. However, similar performance was also given by $U + D + SMART$ feature combination (F1 score=0.708, Accuracy=0.816). When comparing the highest F1 scores of different classifiers, CRF was found to be significantly better than LR but not SP. For the feature combination $U + D + SMART$, the per-class performance using CRF is shown in Table 3, where Support refers to the average number of samples across the 5 folds. We observed that adding other features apart from unigrams, distance, and SMART didn't help to improve performance significantly. In fact, adding other features to SP after unigrams lowered its F1 score. However, when examining the F1 scores for individual classes across all feature combinations, when SMART and distance features were added to unigrams, the models were able to predict the low-frequency classes, especially negotiation, better. The performance on *negotiation* reduces to 0.116 from 0.368 in CRF when the SMART feature is removed from $U + D + SMART$. One can also notice the differences in F1 scores when the SMART tag feature is removed and $U + D + L + T + Se$ is used in Table 2.

SMART Tag Prediction

This task involves classifying each word into one of 11 classes: 10 from the SMART annotation schema plus 'none' for words without any tag. It is similar to a Named Entity Recognition (NER) task, where entities are SMART attributes. We tried both sequential and non-sequential algorithms as many NER tasks are modeled using the former.

We used the five classifiers mentioned in phase prediction: CRF, SP, SVM, LR, and DT and found the same three classifiers CRF, SP, and LR performed the best. We used different combinations of features as listed in Table 4, and report the results in Table 5. The F1 scores are (macro) averaged over

Feature	
W, LW, RW	: Word itself, left word and right word
POS	: Part-of-Speech
LPOS, RPOS	: Left and right word's POS
SNER	: SpaCy Named Entity Recognition
P	: Phases
WE, PWE	: Current and previous word's embeddings

Table 4: List of Features for SMART Tag Prediction

Features	CRF	SP	LR
Baseline	0.57		
W	0.716	0.742	0.420
WE	0.716	0.750	0.500*
W+LW+RW	0.752	0.760	0.734*
W+LW+RW+P	0.752	0.762	0.734
W+LW+RW+WE	0.766*	0.788*	0.758
W+LW+RW+WE+P	0.766	0.790	0.758
W+LW+RW+WE+SNER	0.784	0.790	0.764
W+LW+RW+WE+SNER+P	0.784	0.794	0.768
W+LW+RW+WE+SNER+POS+P	0.778	0.796	0.772
All (from Table 4) except phases	**0.786**	0.796	**0.780**
All (from Table 4)	0.786	**0.802**	0.780

Table 5: SMART tag prediction F1 scores.'*' indicates significant improvement. Boldface indicates highest value in the column

Label	P	R	F1	Support
Activity	0.938	0.946	0.942	122.4
Time	0.724	0.684	0.692	66.0
Location	0.676	0.896	0.722	16.8
Quantity-amount	0.926	0.946	0.934	147.2
Quantity-distance	0.632	0.582	0.552	42.2
Quantity-duration	0.900	0.894	0.882	47.2
Days-name	0.766	0.714	0.728	77.2
Days-number	0.802	0.822	0.810	60.6
Repetition	0.782	0.698	0.722	24.8
Attainability score	0.792	0.708	0.742	13.8
None	0.982	0.988	0.984	5107.2
Macro average	0.808	0.806	0.790	5725.4

Table 6: SMART prediction results per label

all the classes including 'none'. For the baseline, instead of choosing the majority class 'other' which would result in very poor performance, we used a rule-based approach with the help of the existing SpaCy Named Entity Recognizer. It can recognize a variety of named entities, including location, date, time, quantity, cardinal, and many more. We only used the ones that are closest to our SMART attributes. The rules used for each attribute were decided based on the dialogues. E.g., specific activity rule uses the most common activities in the corpus and measurable quantity amount rule is a cardinal POS tag followed by an activity.

We achieved an F1 score of 0.802 over all the categories using all the features in Table 4 with the SP model (when comparing the highest F1 scores of different classifiers, CRF and SP were found to be significantly better than LR). Even though the highest F1 score involves phases as features, they did not provide much improvement overall (only 0.004). The previous word's embedding and POS tag didn't help much either. The F1 scores for individual classes using $W+LW+RW+WE+SNER$ are shown in Table 6. The SP classifier with $W + LW + RW + WE$ combination also had a similar performance. Since SMART tags help in recognition of phases, especially the less frequent ones, we will adopt a pipeline where SMART tags are recognized first, independently of phases, and are then used to recognize phases.

Automatic Goal Extraction

Given the results we just discussed, we chose an SP model for SMART tag prediction with feature combination $W + LW + RW + WE + SNER$, and a CRF model for phase prediction with feature combination $U + D + SMART$ for our goal extraction pipeline. We compare performance on goal extraction, using only SMART tags or adding stages-phases to SMART tags. Our pipeline for goal extraction with the help of phases follows these given steps:

1. Build a SMART prediction model using the training data of k-1 folds and predict SMART tags for the remaining one fold of test data.
2. Using the same k-1 folds, build a phase prediction model. However, when predicting phases for test data, SMART tags predicted in the previous step are used as features along with unigrams and distance. (F1 score = 0.686)
3. Using the model from step 1, extract the last mention for each of the 10 SMART attributes; except
   - in backward-looking goal, for *measurable quantity* (amount, distance and duration) and *measurable days number*, take the last mention only if the message is not in the *follow-up* phase.
   - in forward-looking goal, take the last mention only from (human-annotated) goal setting stage.

The experiments were performed using 5 fold cross-validation over the patients, such that we cover each patient once in the test set. We also created gold standard goal summaries to compare the extracted summary with. For goal extraction baseline, first, the rules for the SMART tag prediction baseline were used to extract the SMART tags and then the last mention for each attribute was taken. We evaluated the models using both existing summary evaluation metrics (BLEU and ROUGE) and accuracy and F-score measures coupled with manual verification.

BLEU and ROUGE results for both forward and backward-looking goals are shown in the first two rows of Table 7(a) and Table 7(b). These metrics suggest that stages-phases do not help in goal extraction. Though BLEU and ROUGE have been widely used for fast and easy summary evaluation, they are only sensitive to exact word match. That means, it doesn't matter if a given word, say 'two', is classified as *number of days* or *measurable distance*, BLEU and ROUGE will output a high score as long as 'two' is a part of the reference summary. Hence, we will look at the per attribute results and show that the results contrast with the *(a) forward-looking goal*

Evaluation	SMART tags	SMART + stages-phases
BLEU (unigrams)	0.47	0.48
ROUGE (unigrams)	0.62	0.62
slot-value exists (Avg. Fscore)	0.77	0.88
Type match (Avg. Accuracy)	0.97	0.98
Partial match (Avg. Accuracy)	0.79	0.85
Complete match (Avg. Accuracy)	0.75	0.81

(b) backward-looking goal

Evaluation	SMART tags	SMART + stages-phases
BLEU (unigrams)	0.48	0.48
ROUGE (unigrams)	0.63	0.65
slot-value exists (Avg. Fscore)	0.77	0.80
Type match (Avg. Accuracy)	0.95	0.96
Partial match (Avg. Accuracy)	0.77	0.81
Complete match (Avg. Accuracy)	0.73	0.77

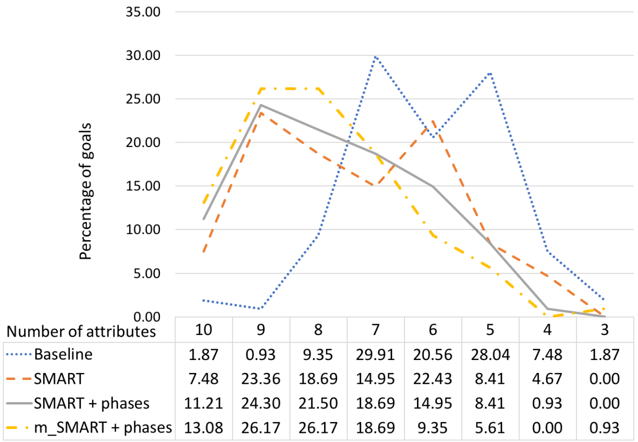
Table 7: Goal extraction evaluation: BLEU and ROUGE scores average over summaries; F-scores and accuracies average over slots.

BLEU and ROUGE scores.

We evaluated the extracted slot-values against the gold-standard slot-values at multiple levels:

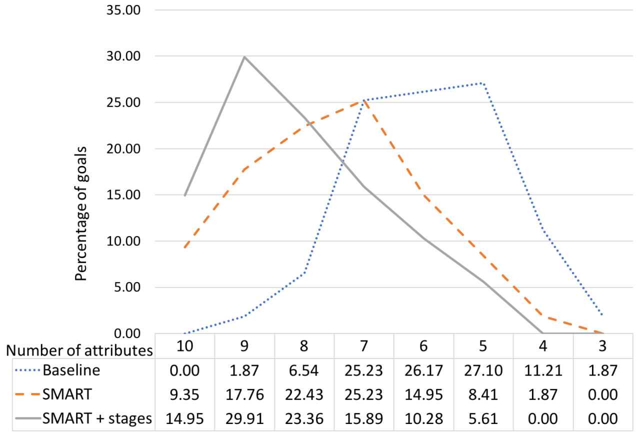
1. Attribute presence test: checks if the algorithm correctly predicts the existence or non-existence of a slot-value
2. Attribute type match: checks if the extracted value belongs to the correct attribute category. E.g., ‘20 minutes’ under *specific time* will be considered a type mismatch as it should be *measurable amount duration*.
3. Partial attribute match: checks if the extracted value at least covers part of the gold standard attribute. E.g., ‘Wednesday’ extracted instead of ‘Monday, Wednesday’ will be counted as a partial match, but ‘blocks’ instead of ‘5 blocks’ is not a partial match but a type match.
4. Complete attribute match: checks if the extracted value exactly matches the gold standard attribute except for punctuation, spaces and the measuring unit (only if implicit through the attribute type). E.g., ‘2 days’ and ‘2’ under *measurable days number* will be considered a complete match, however ‘2 blocks’ and ‘2’ under *measurable amount distance* will not be considered a complete match.
5. Complete goal match: checks if all the 10 slot values match completely for a given goal.

The results for the first four are shown in Table 7(a) for forward-looking goals and Table 7(b) for backward-looking goals. SMART plus stages-phases performed better than only SMART tags at all four levels. The performance over the entire goal is shown in Figures 2 and 3 for backward- and forward-looking goals respectively. The graphs show the performance over the different number of attributes, where 10 means that all the 10 attributes were correct for a given goal. Percentage of goals with less than 3 attributes correct is 0%. Since the number of attributes are in descending order in the graph, the higher the percentage on the left, the better.



Number of attributes	10	9	8	7	6	5	4	3
Baseline	1.87	0.93	9.35	29.91	20.56	28.04	7.48	1.87
SMART	7.48	23.36	18.69	14.95	22.43	8.41	4.67	0.00
SMART + phases	11.21	24.30	21.50	18.69	14.95	8.41	0.93	0.00
m_SMART + phases	13.08	26.17	26.17	18.69	9.35	5.61	0.00	0.93

Figure 2: Percentage of goals with given number of attributes correct for backward-looking goals



Number of attributes	10	9	8	7	6	5	4	3
Baseline	0.00	1.87	6.54	25.23	26.17	27.10	11.21	1.87
SMART	9.35	17.76	22.43	25.23	14.95	8.41	1.87	0.00
SMART + stages	14.95	29.91	23.36	15.89	10.28	5.61	0.00	0.00

Figure 3: Percentage of goals with given number of attributes correct for forward-looking goals

For now only consider the first three rows in Figure 2; we will discuss m_SMART later. For both backward and forward looking-goals, the best performance is achieved using ‘SMART tags + phases’ and ‘SMART tags + stages’ respectively. For backward-looking goals, we obtained 11% of the goals with all the 10 attributes correct and 57% of goals with at least 8 attributes correct (adding percentages for 10, 9 and 8 attributes correct). Similarly, for forward-looking goals, we obtained 15% of goals with all the 10 attributes correct and 68% of goals with at least 8 attributes correct.

When we analyzed the per attribute performance, *measurable quantity amount* and *measurable days name* performed the worst. One of the reasons for the poor performance of *measurable quantity amount* was the recognition of the amounts related to the current progress. E.g., if the goal was 7000 steps, and the coach sends the message ‘*5000 steps* done, *2000 steps* more’, we don’t want the model to recognize the intermediate progress. Since we had also annotated SMART tags as *accomplished, remaining, previous* or *other*, we decided to leverage that and treat them as negative examples for the SMART prediction task (the first step in goal extraction). That is, the words with those features were treated as having 'none' tag. This will reduce the recognition of such words and hence should improve the performance. We call it modified SMART prediction model (m_SMART).

We saw an improvement of almost 18% when looking at the complete attribute match alone for *measurable quantity amount*. This is also reflected in the final goal extraction performance shown in Figure 2 indicated with 'm_SMART + phases'. The maximum percentage of 13% goals with all 10 attributes correct was obtained using the modified SMART model, which is a 2% increase from the previous SMART model. When we look further at the percentage of goals with at least 8 attributes correct, we see that 65% of goals had at least 8 attributes correct when using the modified SMART prediction model as compared to 57% before. However, we achieved BLEU and ROUGE scores of 0.44 and 0.56 for the modified pipeline, which is worse than the original pipeline performance of 0.48 and 0.65, and confirms that these two metrics are not the best fit for our type of summaries.

Conclusions and Future Work

In this paper, we discussed our work towards building a virtual assistant health coach that can help patients to live a more active lifestyle. We focused on the NLU component and modeled the goal summarization pipeline using slot values and higher-level conversation flow. The current pipeline does not take any utterance level intent into account while generating the goal summary. Therefore, our next step is to improve the goal summarization results by incorporating dialogue acts at the utterance level. We will also run our first human evaluation of the goal summarization pipeline in our next round of data collection. We plan to provide the current week's goal summary to health coaches in the application interface, where they will be able to provide binary feedback of correct/incorrect for individual summaries.

Acknowledgements

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

References

Althoff, T.; Clark, K.; and Leskovec, J. 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.

Andrade, A. S.; McGruder, H. F.; Wu, A. W.; Celano, S. A.; Skolasky Jr, R. L.; Selnes, O. A.; Huang, I.-C.; and McArthur, J. C. 2005. A programmable prompting device improves adherence to highly active antiretroviral therapy in hiv-infected subjects with memory impairment. *Clinical Infectious Diseases* 41(6):875–882.

Bickmore, T.; Asadi, R.; Ehyaei, A.; Fell, H.; Henault, L.; Intille, S.; Quintiliani, L.; Shamekhi, A.; Trinh, H.; Waite, K.; et al. 2015. Context-awareness in a persistent hospital companion agent. In *International Conference on Intelligent Virtual Agents*. Springer.

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

Blackwell, D., and Clarke, T. 2018. State variation in meeting the 2008 federal guidelines for both aerobic and muscle-strengthening activities through leisure-time physical activity among adults aged 18-64: United states, 2010-2015. *National Health Statistics Reports* (112):1–22.

Booth, F. W.; Roberts, C. K.; Thyfault, J. P.; Ruegsegger, G. N.; and Toedebusch, R. G. 2017. Role of inactivity in chronic diseases: evolutionary insight and pathophysiological mechanisms. *Physiological Reviews* 97(4):1351–1402.

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

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

Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 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).

Gupta, I.; Di Eugenio, B.; Ziebart, B.; Liu, B.; Gerber, B.; Sharp, L.; Davis, R.; and Baiju, A. 2018. Towards building a virtual assistant health coach. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 419–421. IEEE.

Gupta, I.; Di Eugenio, B.; Ziebart, B.; Liu, B.; Gerber, B.; and Sharp, L. 2019. Modeling health coaching dialogues for behavioral goal extraction. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1188–1190. IEEE.

Kivelä, K.; Elo, S.; Kyngäs, H.; and Kääriäinen, M. 2014. The effects of health coaching on adult patients with chronic diseases: a systematic review. *Patient Education and Counseling* 97(2).

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81. Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311–318.

Pérez-Rosas, V.; Mihalcea, R.; Resnicow, K.; Singh, S.; Ann, L.; Goggin, K. J.; and Catley, D. 2017. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 1128–1137.

Pérez-Rosas, V.; Sun, X.; Li, C.; Wang, Y.; Resnicow, K.; and Mihalcea, R. 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)*.

Piercy, K. L.; Troiano, R. P.; Ballard, R. M.; Carlson, S. A.; Fulton, J. E.; Galuska, D. A.; George, S. M.; and Olson, R. D. 2018. The physical activity guidelines for americans. *JAMA* 320(19).

Shamekhi, A.; Bickmore, T.; Lestoquoy, A.; and Gardiner, P. 2017. Augmenting group medical visits with conversational agents for stress management behavior change. In *International Conference on Persuasive Technology*, 55–67. Springer.

Song, H.; May, A.; Vaidhyanathan, V.; Cramer, E. M.; Owais, R. W.; and McRoy, S. 2013. A two-way text-messaging system answering health questions for low-income pregnant women. *Patient Education and Counseling* 92(2):182–187.

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