Modeling Health Coaching Dialogues for Behavioral Goal Extraction

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

***Abstract*—In this paper, we will discuss our framework for summarizing goals discussed during health coaching dialogues. This can help coaches to recall patients' goals without reading the conversations. We build two supervised classification models, one for extracting the slot-values (goal attributes) and another to model the dialogue flow (stages-phases) of the conversation. Using these two models and heuristics, we build our goal extraction pipeline.**

***Index Terms*—health coaching, virtual assistant, goal extraction, SMART goals**

I. Introduction

Health coaching (HC) has been identified as a successful method for motivating and maintaining health behavior changes. Unfortunately, personal HC is time- and resource-intensive, and cannot scale up. Several influential papers focus on developing conversational systems that can provide automated coaching to patients [1], [2]. But 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 implementation. Therefore, we aim to build a virtual coach that can help patients set and achieve a Specific, Measurable, Attainable, Realistic and Time-bound (S.M.A.R.T.) goal [3] via SMS.

A typical text-based Dialogue System (DS) consists of 3 components: Natural Language Understanding (NLU) module, Dialogue Manager, and Natural Language Generator. My work focuses on the NLU module and uses it to summarize/extract patients' goals. This can assist human health coaches to recall the goals without reading the conversations. An NLU module consists of recognizing task-specific slots and the user's intent. In our data, slots are the SMART goal attributes and intents are the higher-level stages-phases shaping the conversations. We believe that recognizing the stages-phases in the conversation such as negotiation and discussion of barriers can help improve the performance of goal extraction [4]. These stages and phases are more abstract, but otherwise analogous to tasks and sub-tasks as defined in task-oriented dialogue systems [5].

Funded by the National Science Foundation through SMART and Connect Health (SCH) and EArly-concept Grants for Exploratory Research (EAGER).

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

In this paper, we present two supervised classification models: one for predicting phases and the second for predicting goal attributes and evaluate the importance of one in predicting the other. We then use these models for goal extraction.

II. Related work

Our research is relevant to modeling both health dialogues and more in general, dialogue structure.

Dialogues in Health Domains. The potential for automated systems to provide health coaching has attracted significant attention in the past three decades. At one end of a vast spectrum of work lie simple, pre-programmed, reminder-based systems used in interventions such as medication adherence or adopting healthy habits [6], [7]. At the other end of the spectrum, conversational agents fully interact with users and help them manage stress and assist during hospital visits [8], [9]. However, the majority of these systems provide a predefined set of input, use "dialogue recipes" to represent the dialogue flow, and use templates to generate output. Work by Althoff, Clark, and Leskovec [10] and by Pérez-Rosas et al. [11] respectively deal with issues such as mental health and motivational interviewing, but not goal setting.

Modeling Dialogue Structure is a crucial step in a dialogue system. Often Dialogue Acts (DAs) are used to model the intentions of speakers, and their sequencing is recognized via Hidden Markov Models (HMM) or vector-based classifiers. Alternatively, unsupervised learning can be used for learning task structure in the dialogue [12], [13].

We believe we are the first ones to have collected health coaching conversations with SMART goal setting and analyzed their structure [3]. The SMART approach is over three decades old and has been rigorously adopted to set realistic and manageable goals in different fields. It has been shown that goal setting and action planning helps patients adopt healthy behaviors and manage chronic diseases [4]. While the structure of such dialogues derives from techniques such as motivational interviewing [11], to the best of our knowledge no computational treatment of this sort of data exists.

III. Data Collection and Annotations

We recruited 28 patients (21-65 years) and a health coach, who conversed with these patients via SMS for about a month (4 weeks) and helped them set a new SMART goal every week

TABLE I
COUNTS FOR SMART TAGS IN THE CORPUS (κ >0.69 FOR EACH TAG)

Tag	Feature	Slot Value
Specificity	Activity	671
	Time	131
	Location	41
Measurability	Quantity	627
	Days	303
	Repetition	69
Attainability		70

TABLE II
COUNTS FOR STAGE-PHASE TAGS IN THE CORPUS (κ = 0.93 TOGETHER)

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 Action	Refining	16	4
	Anticipate Barrier	8	4
	Solve Barrier	25	7
	Negotiation	23	6
	Follow up	1348	120

based on their past week's performance. The patients were given a Fitbit Alta and the coach used the Fitbit application to monitor patients' progress. The coach also sent reminders, negotiated goals, and provided motivational feedback. Only one patient didn't finish the study due to health reasons. We have a corpus of 2853 messages, to which patients and the coach contributed almost equally.[1] Even if the number of messages decreases from week 1 (33 on average) to week 4 (21.74), numerous messages were shared every week. We also designed our schemas and annotated the dialogues for SMART attributes to capture slot values and coaching stages-phases to capture dialogue flow [15]. Two annotators labeled the data and reliability was calculated using the kappa coefficient [14]. Table I & II show the counts for each tag and kappa scores.

IV. GOAL SUMMARIZATION PIPELINE

During goal-setting, the patient and the coach would collaboratively negotiate a realistic goal. However, the patients sometimes need to change their goals during the week due to unseen circumstances or difficulty in accomplishing the goal. This makes the information about the goal to be distributed throughout the dialogue. We hypothesize that understanding the current message's stage and phase can help to identify goal modifications and extract the final modified goal. In total there are 107 goals, 4 per patient (one per week) with one exception where the patient took longer to set one of the goals and therefore, only had 3 goals in 4 weeks.

For a benchmark, we used human-annotated SMART tags and extracted the last mention of all the attributes at the end of the week. We then compared it against the human-generated gold standard for goals. This resulted in an accuracy of 22.43% i.e. only 22.43% of the goals were extracted completely correct. However, when stages-phases were added, the accuracy increased to 40.19%. This shows that stages-phases do help in goal extraction.

[1]We cannot share the data due to human subject protection (cf. the US Health Insurance Portability and Accountability Act (HIPAA)). We have collected a larger corpus recently, and asked subjects for permission to share their de-identified conversations.

TABLE III
PHASE PREDICTION RESULTS PER LABEL

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 IV
SMART PREDICTION RESULTS PER LABEL

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

Hence, we decided to model the problem of goal extraction in three parts: (1) predicting the current phase (2) predicting SMART tag attributes and (3) using the models from 1 and 2 to extract the goal. We will also conclude that we should infer phases based on SMART tags, and not vice-versa.

Phase Prediction. Only 39 unique transitions occur out of 111 possibilities in our data set (10 unique stage-phase categories plus the beginning and end of the week). Thus, we decided to try both sequential and non-sequential classification algorithms for predicting phases. For sequential algorithms, we modeled a set of messages in one week as one sequence.

We divided data into train (80%) and test (20%) and performed 5-fold cross-validation. We used supervised classification models: Conditional Random Fields (CRF), Structured Perceptron (SP), Support Vector Machines (SVM), Logistic Regression (LR) and Decision Trees (DT). We tried different combination of features: unigrams (U), message distance from the top in a week (D), presence/absence of each SMART attribute (SMART), sentence length (L), normalized time difference between messages (T), sender (Se) and Google word embedding (WE). The F1 score of 0.708 was achieved with $U+D+SMART$ features using CRF. Per label performance is shown in Table III. When SMART and distance features were added to unigrams, the models were able to predict the low-

frequency classes, especially negotiation, much better. The F1 score on *negotiation* reduces to 0.116 in CRF without SMART.

SMART Prediction involves classifying each word into one of the 11 classes shown in Table IV where 'none' is for words without a tag. It is similar to a Named Entity Recognition (NER) task, where entities for us are SMART attributes. We tried both sequential and non-sequential algorithms as many NER tasks are modeled using the former.

We used the same five classifiers mentioned earlier and found SP performed the best. We used different combinations of features: word, left word and right word (W, LW, RW), part-of-speech tags of the words (POS, LPOS, RPOS), phases (P), Google word embedding (WE), and SpaCy NER (SNER) output. We achieved an F1 score of 0.790 over all the categories using (W+LW+RW+WE+SNER) feature combination with the SP model. Per label F1 scores are shown in Table IV. When comparing the highest F1 scores, CRF and SP performed significantly better than other classifiers in both phase and SMART prediction tasks. Phases as a feature did not provide much improvement in the results. Since SMART tags help 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.

Goal Extraction. We chose SP model for SMART tag prediction with $W+LW+RW+WE+SNER$ features and CRF model for phase prediction with $U+D+SMART$ features for goal extraction pipeline. We performed goal extraction using only SMART tags and also using SMART tags plus stages-phases. In former, we first predicted all the SMART tags and extracted the last mention for each of the 10 attributes. This resulted in an accuracy of 7.48%, where all the 10 attributes matched with the gold standard. For SMART tags plus phases, we extracted the last mention for each of the 10 SMART attributes; except for *measurable quantity* (amount, distance and duration) and *measurable days number*, we took the last mention only if the message was not in *follow-up* phase. This gave an accuracy of 13.08%. (Note: Benchmark accuracy is the maximum we can achieve given our current pipeline.)

Though the accuracy is low, most of these errors are due to the goals (36%) which involve at least one SMART slot distributed over multiple messages or phrases in the message. E.g., if a patient's goal was initially for Monday and Wednesday, but then the patient says 'I would like to add Tuesday too', the system needs to identify that the goal has changed to 'Monday, Tuesday and Wednesday' and not to 'Tuesday' only. However, our current pipeline replaces the slot value with the new value. Moreover, we consider a goal to be correct if all the 10 attributes match, which is a very strict measure. We had 26.17% of goals with 9 attributes correct and another 26.17% with 8 attributes correct. That means over 65% of extracted goals had at least 8 correct attributes. We also evaluated our results using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [16], a well known metric for summarization and machine translation and achieved an F-score of 0.57 using ROUGE over unigrams. Though it's not the most efficient metric for comparison, it is one of the standard metrics that is still being used for evaluation.

V. Conclusion and Future Work

We built a goal summarization pipeline to help health coaches to recall the patients' goals. Given the complex decision-making nature of our health coaching dialogue, we plan to extend our pipeline and incorporate dialogue act annotations on utterance or message level to understand the sender's intent. We believe understanding if a given message is a suggestion, rejection, modification can help to improve goal summarization. Also, we plan to evaluate our pipeline on the newly collected data and use it for the next round of data collection for extrinsic evaluation.

References

[1] 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), p.e1.

[2] Bickmore, T.W., Kimani, E., Trinh, H., Pusateri, A., Paasche-Orlow, M.K. and Magnani, J.W., 2018, November. Managing Chronic Conditions with a Smartphone-based Conversational Virtual Agent. In IVA (pp. 119-124).

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

[4] Bodenheimer, T., Davis, C., and Holman, H., 2007. Helping patients adopt healthier behaviors. Clinical Diabetes, 25(2):66–70.

[5] Chotimongkol, A., 2008. Learning the structure of task-oriented conversations from the corpus of in-domain dialogs. Ph.D. thesis, Carnegie Mellon University, Language Technologies Institute.

[6] Schollmeyer, T. A., and Elsmore, J. C. (1985). U.S. Patent No. 4,504,153. Washington, DC: U.S. Patent and Trademark Office.

[7] Riva, A., Smigelski, C. and Friedman, R., 2000. WebDietAID: an interactive Web-based nutritional counselor. In Proceedings of the AMIA Symposium (p. 709). American Medical Informatics Association.

[8] Shamekhi, A., Bickmore, T., Lestoquoy, A. and Gardiner, P., 2017, April. Augmenting group medical visits with conversational agents for stress management behavior change. In International Conference on Persuasive Technology (pp. 55-67). Springer, Cham.

[9] Bickmore, T. et al., 2015, August. Context-awareness in a persistent hospital companion agent. In International Conference on Intelligent Virtual Agents (pp. 332-342). Springer, Cham.

[10] 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, pp.463-476.

[11] Pérez-Rosas, V. et al., 2017, April. Predicting counselor behaviors in motivational interviewing encounters. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers (pp. 1128-1137).

[12] Ritter, A., Cherry, C. and Dolan, B., 2010, June. Unsupervised modeling of twitter conversations. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (pp. 172-180). Association for Computational Linguistics.

[13] Purver, M., Griffiths, T.L., Körding, K.P., and Tenenbaum, J.B., 2006, July. Unsupervised topic modelling for multi-party spoken discourse. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (pp. 17-24). Association for Computational Linguistics.

[14] Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), pp.37-46.

[15] Gupta, I. et al., 2018. Towards building a virtual assistant health coach. In 2018 IEEE International Conference on Healthcare Informatics (ICHI), 419–421. IEEE.

[16] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 7481, 2004.