Summarizing Behavioral Change Goals from SMS Exchanges to Support Health Coaches

Itika Gupta,[1] **Barbara Di Eugenio,**[1] **Brian D. Ziebart,**[1] **Bing Liu,**[1]
Ben S. Gerber,[2] **Lisa K. Sharp**[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, liub}@uic.edu
{bgerber, sharpl}@uic.edu

Abstract

Regular physical activity is associated with a reduced risk of chronic diseases such as type 2 diabetes and improved mental well-being. Yet, more than half of the US population is insufficiently active. Health coaching has been successful in promoting healthy behaviors. In this paper, we present our work towards assisting health coaches by extracting the physical activity goal the user and coach negotiate via text messages. We show that information captured by dialogue acts can help to improve the goal extraction results. We employ both traditional and transformer-based machine learning models for dialogue acts prediction and find them statistically indistinguishable in performance on our health coaching dataset. Moreover, we discuss the feedback provided by the health coaches when evaluating the correctness of the extracted goal summaries. This work is a step towards building a virtual assistant health coach to promote a healthy lifestyle.

1 Introduction

Physical activity (PA) is extremely beneficial to one's health, as it reduces the risk for serious health problems like type 2 diabetes and heart diseases, and also helps to improve mood and reduce depression and anxiety (Manley, 1996; Stephens, 1988). Yet, only 45% of the US adult population met the federal guidelines for PA in 2016 (Piercy et al., 2018). Such findings suggest that a majority of people are unable or not motivated enough to engage in PA (Teixeira et al., 2012).

Health coaching has been identified as a successful method for facilitating health behavior changes: a professional provides evidence-based interventions, support for setting realistic goals and encouragement for goal adherence (Palmer et al., 2003; Kivelä et al., 2014). Health coaching has its origin in motivational interviewing (MI) and is guided by the principle that patients need to identify their own motivation to be successful in achieving health behavior changes (Miller and Rollnick, 2002; Huffman, 2009). Unfortunately, personal health coaching is expensive, time-intensive, and may have limited reach because of distance and access.

As a consequence, researchers have been exploring the use of technology such as computers and (mobile) phones (McBride and Rimer, 1999; Krebs et al., 2010; Free et al., 2013; Buhi et al., 2013) to promote health behavior changes for a while now. Mobile health technologies (mHealth) have been particularly effective due to their accessibility and ability to reach large populations at low cost. Currently, about 96% of the US population owns a cellphone and 81% owns a smartphone (Sheet, 2018). Therefore, we aim to build a dialogue-based virtual assistant health coach to help patients set physical activity goals via text messages (SMS) (Doran, 1981). Studies have shown that setting specific and challenging goals leads to better performance than setting abstract or easy goals (Locke and Latham, 2002; Bodenheimer et al., 2007). SMART (Specific, Measurable, Attainable, Realistic, and Timebound) is one such goal setting approach that helps to create specific, measurable, and manageable goals and provides a clear path to success.

In this paper, we focus on the natural language understanding (NLU) module of the dialogue system, grounded in two health coaching datasets we collected, and show its application to goal summarization. These goal summaries will help the health coaches to easily recall patients' past goals and use them to suggest a realistic future goal. Currently, our health coaches use external documents like Microsoft Excel to keep track of patients' goals. We conducted an evaluation with the health coaches where they assessed the correctness and usefulness of automatically generated goal summaries in an offline setting. In our future work, health coaches

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 276–289
July 29–31, 2021.

will use the goal summarization module during real-time health coaching and we will evaluate its usefulness in the real world. The main contributions of this paper are:

- We propose a two-step goal extraction process that uses phases and dialogue acts to identify the correct goal attributes and show that dialogue acts help more than phases.
- We employ both traditional and transformer-based machine learning models for dialogue act prediction and find them statistically indistinguishable in performance.
- We evaluate the correctness and usefulness of the goal summaries generated by our model from the health coaches' perspective.

2 Related Work

Dialogue agents in healthcare. Researchers have explored the use of technology to extend the benefits of counseling to millions of people who can't access it otherwise. Pre-programmed text messages were used by Bauer et al. (2003) to send responses based on patients weekly reporting of their bulimic symptomatology. More engaging systems involve a back and forth conversation even if it is solely based on keyword matching (Weizenbaum, 1966). These conversations are sometimes made more human-like with the help of an animated character that uses both verbal and non-verbal cues (Shamekhi et al., 2017; Bickmore et al., 2018). However, these systems need to be installed as a separate application and require a smartphone. In contrast, text messages are low cost, and afford a 'push' technology that allows both the user and the agent to initiate a conversation, and that requires no extra effort such as installation or logging in (Aguilera and Muñoz, 2011).

Some of the recent dialogue-based assistants in the field include *Woebot*, a commercialized dialogue agent that helps young adults with symptoms of depression and anxiety using cognitive behavior theory (Fitzpatrick et al., 2017). *Woebot* accepts natural language input and uses a decision tree to decide the response. *Vik Asthma* is another commercialized dialogue system that is designed to remind patients to take their medications and answer questions about asthma (Chaix et al., 2020). *NutriWalking* application helps sedentary individuals with regular exercise (Mohan et al., 2020). It consists of multiple choice options for the user to choose from and relies on user reporting their progress rather than using input from activity trackers like Fitbits. Kocielnik et al. (2018) used Fitbit and SMS to build a *Reflection Companion* that allows users to reflect on their PA performance with a series of follow-up questions, however, no goal-setting is involved.[1]

Interactions in these dialogue agents are still mostly scripted. Dynamic interactions require large datasets that are unfortunately scarce in the health domain due to privacy reasons. Moreover, collecting and labeling data particularly in real scenarios is resource intensive. This limits the researchers from applying state-of-the-art deep learning techniques and end-to-end approaches for building dialogue agents that require large datasets. Researchers like Althoff et al. (2016) and Zhang and Danescu-Niculescu-Mizil (2020) were able to access a large counseling conversations dataset from the Crisis Text Line (CTL), a free 24/7 crisis counseling platform for a mental health crisis, for computational analysis through a fellowship program with CTL. Online sources such as Reddit have also been used for analyzing empathy in conversations, but consist of question-answer pairs and not dialogues (Sharma et al., 2020). Lastly, Shen et al. (2020) used the MI dataset collected by Pérez-Rosas et al. (2016) to build a model that can generate sample responses of type *reflection* to assist counselors. As far as we know, no existing work has focused on building a dialogue agent involving coaching components such as negotiation and feedback for promoting PA using SMART goal setting.

Dialogue act (DA) modeling. This task involves finding the intent behind the speaker's utterance in a dialogue such as *request*, *clarification*, and *acknowledgment*. The DA tags may differ depending on the dialogue's domain. E.g., negotiation dialogues might involve tags like *offer*, *accept*, and *suggest*. As a result, numerous DA schemas have emerged over time (Core and Allen, 1997; Bunt, 2009; El Asri et al., 2017; Budzianowski et al., 2018). However, the majority of them are difficult to reuse due to their complexity and lack of generalizability to other domains.

Efforts have been devoted to create a standardized schema that can be used for multiple datasets in different domains. One such effort led to the formation of the ISO 24617-2, the international ISO standard for DA annotations (Bunt et al., 2010). It provides a domain- and task-independent DA

[1]Many studies show Fitbit can help increase physical activity (Ringeval et al., 2020), but here we are interested in approaches with dialogue capabilities.

schema with 56 DAs organized into nine dimensions. Paul et al. (2019) proposed a universal DA schema by aligning tags from different datasets such as the Dialogue State Tracking Challenge 2 (Henderson et al., 2014), Google Simulated Dialogue (Shah et al., 2018), and MultiWOZ 2.0 (Budzianowski et al., 2018) together. Mezza et al. (2018) reduced the ISO schema to 10 DAs and showed their applicability to datasets like Switchboard (Leech and Weisser, 2003), MapTask (Anderson et al., 1991), and VerbMobil (Alexandersson et al., 1998). On account of not reinventing the wheel, we used the ISO schema for our dialogues (Bunt et al., 2017a). Since many of the DAs didn't apply to our dataset such as *turn take/grab, stalling*, and *pausing*, we reduced the schema to only 12 DAs, mostly following Mezza et al. (2018).

Early work for DA modeling involved treating the task as a structured prediction or text classification problem. Stolcke et al. (2000) used Hidden Markov Models (HMM) to model the dialogue structure, where individual DAs were treated as observations and n-grams were used to model the probability of the DA sequence. They also used acoustic correlates of prosody as raw features in the HMM model. Researchers have also explored non-verbal cues such as body postures to better understand a user's intent during a tutorial dialogue (Ha et al., 2012). Since then, deep learning techniques have also been applied to the task (Kumar et al., 2020; Anikina and Kruijff-Korbayova, 2019). Convolutional Neural Networks (CNN) were also used for intent classification of a query (Hashemi et al., 2016). However, queries can be treated as individual sentences without any context. Given context is important in a dialogue, we experiment with approaches that can take dialogue history into account such as Conditional Random Fields (CRF) (Lafferty et al., 2001) and recent transformer-based BERT (Bidirectional Encoder Representations from Transformers) models (Devlin et al., 2019). In particular, we use the work by Wu et al. (2020) and Cohan et al. (2019) as the guide for our BERT-based DA prediction models.

3 Datasets and Annotations

No health coaching dialogue dataset is publicly available. Therefore, to understand the feasibility of using SMS for health coaching, the challenges with patient recruitment and retention, and conversation flow between the coaches and the patients, we collected two health coaching datasets (Dataset 1 and Dataset 2; Dataset 2 is available upon request[2], while Dataset 1 cannot be shared due to lack of subject consent). To collect Dataset 1, we hired one health coach who coached 28 patients, recruited at one of UI Health's internal medicine clinics, for 4 weeks (since one patient didn't finish the study, we exclude their data). The health coach, trained in SMART goal setting, helped patients to set a new SMART physical activity goal every week. The health coach used Mytapp, a web-based application developed by one of our collaborators and validated in other text-based health monitoring studies (Stolley et al., 2015; Kitsiou et al., 2017), to text the patients, who used their smartphones' texting service. The patients were also given a Fitbit to track their progress and the coach could access patients' Fitbit data on Mytapp.

The data collection process was similar for Dataset 2, except we hired three new health coaches and 30 different patients, and doubled the duration to 8 weeks. Since one patient lost their Fitbit and one almost stopped responding after 2 weeks, we only consider 28 patients' data for Dataset 2. Dataset 1 comprises 2853 text messages (54% coach, 46% patients) and Dataset 2 comprises 4134 text messages (58% coaches, 42% patients). In Dataset 1, the average number of words per message for coach is 13.74 ± 9.76 and for patient is 7.68 ± 8.19, while in Dataset 2, the average number of words per message for coach is 19.27 ± 10.37 and for patient is 9.28 ± 10.74.

All our models in this paper are built on Dataset 1 as it was collected first and hence, we only had gold standard annotations for that. Dataset 2 was collected two years later. We did however annotate a subset of Dataset 2 to evaluate the performance of our models on it. More information on the two datasets is available in Gupta et al. (2020a).

We performed a three-tier annotation on Dataset 1: (1) stages and phases (2) dialogue acts, and (3) SMART attributes. An example annotated with all the three schemas is shown in Figure 1. The stages and phases schema captures the higher-level conversation structure and consists of two stages: *goal setting* and *goal implementation*. The *goal setting* stage consists of five phases: *identification*, *refining*, *negotiation*, *solve barrier*, and *anticipate barrier*. The *goal implementation* consists of the

[2]Dataset 2 is available upon request because subjects consented to share their data but did not explicitly consent to make it available on the web.

Stage: Goal Setting

Phase: Goal Identification
Coach: Think about this week, let's call it week 1 of 4. [***Directive***] Now what goal could you make that would allow you to do more walking (***Specific activity***)? [***Set question***]
Patient: I can take the stairs (***Specific activity***) at work (***Specific location***) for the work week. [***answer***]

Phase: Goal Refining
Coach: Sounds good. [***Feedback***] So will this be for coming, going and at lunch time (***Specific time***)? [***Propositional question***]
Patient: I will do at least twice (***Measurable repetition***) during workdays (***Measurable days name***). [***Answer***]

Phase: Anticipate Barrier
Coach: What might get in the way of you accomplishing your goal? [***Set question***]
Patient: Well If im pressed for time that could stop me. [***Answer***] But i think i can fit in at least twice a day (***Measurable repetition***). [***Inform***]
Coach: On a scale of 1-10 with 10 being very sure and 1 not at all sure. [***Directive***] How sure are you that you will accomplish your goal? [***Set question***]
Patient: 9 (***Attainability score***) [***Answer***]

Phase: Solve Barrier
Coach: That is pretty sure but I want you to succeed... [***Inform***] so how can you make that a 10? [***Set question***]
Patient: Well, its my first. Lol. Im not exactly an exercise pro, but with the coaching help im sure i can have that at 10 by next week. [***Answer***]

Stage: Goal Implementation

Phase: Follow up
Coach: Good morning! [***Salutation***] How is your goal for this week going so far? [***Set question***]
Patient: Going great. [***Answer***]

Figure 1: A dialogue excerpt annotated with stages-phases, dialogue acts, and SMART attributes schemas

same phases, minus the *identification* phase, and plus an additional *follow up* phase. DAs capture the general intent of the sender's message at the utterance level (a message can contain one or more utterances). We use a set of 12 tags: *set question*, *choice question*, *propositional question*, *inform*, *answer*, *commissive*, *directive*, *feedback*, *apology*, *salutation*, *thanking*, and *self correction*. This is the same set of tags used by Mezza et al. (2018), except we added the *answer* and *self correction* tags from the original ISO standard schema. This is because it is important for us to differentiate between *inform*, *answer*, and *self correction* tags for the goal summarization pipeline. The SMART attributes schema captures the domain-specific slot values at the word-level and consists of 10 attributes: *specific activity*, *specific time*, *specific location*; *measurable quantity amount*, *measurable quantity distance*, *measurable quantity duration*, *measurable days name*, *measurable days number*, *measurable repetition*; and *attainability score*. To measure intercoder agreement, two annotators annotated four patients' data (447 messages) and obtained an excellent $\kappa = 0.93$ for phases; for SMART attributes, κ ranges between ≈ 0.5 for *Attainability* to ≈ 0.9 for *Specificity* and *Measurability*.[3]

[3]We didn't calculate kappa for dialogue acts as this schema has been validated on many other datasets (Bunt et al., 2017b).

4 Goal Extraction Approach

It is usually assumed that users have a specific goal in mind when interacting with a goal-oriented dialogue system. As the user attempts to complete one sub-task after another in order to achieve the final goal, the dialogue becomes easy and sequential. However, some use-cases involve a decision-making process. E.g., when booking a flight ticket, the user might want to compare prices for different days, times, destinations, etc. In such cases, the system must keep all options available instead of simply replacing the slots in order. Similarly, in our dataset, we noticed complex decision-making behavior where different entities are introduced by both the coach and the patient, some of these entities are then accepted, others rejected or forgotten. The conversation also consists of various SMART attribute values that refer to the patient's current progress towards the goal. Hence, the coaches need to scroll back through the patient conversations to recall the original goal and determine if the goal was met. A goal summary readily available for each patient can save time for health coaches and provide an idea for a realistic future goal. The correct goal summary for the conversation in Figure 1 will be -

activity: 'stairs', ***location***: 'at work', ***days name***: 'workdays', ***repetition***: 'twice a day', ***score***: '9'

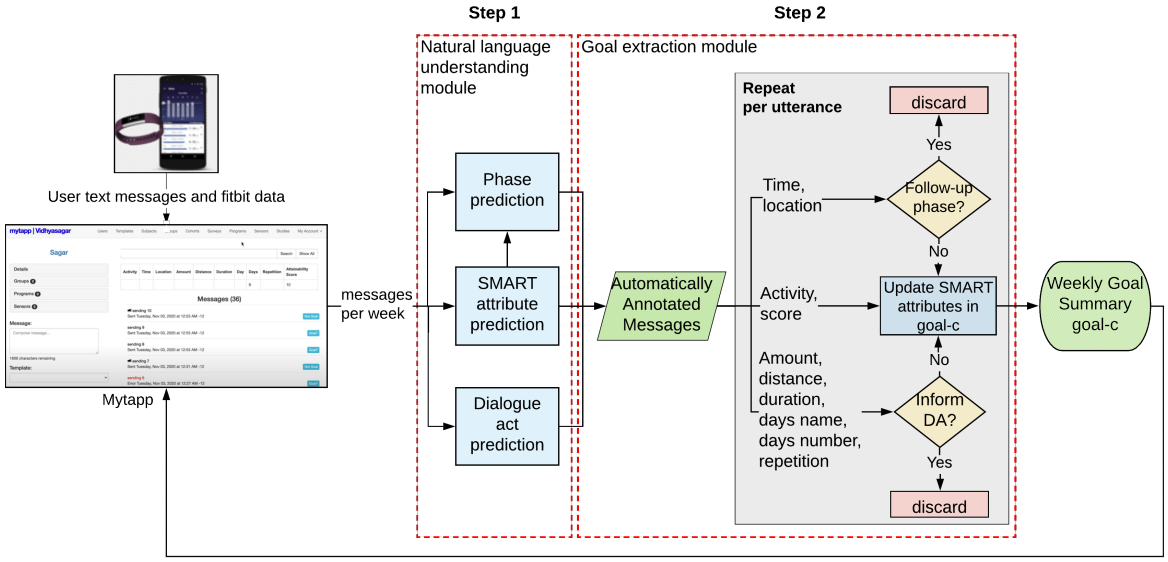

Figure 2: Goal extraction architecture

Figure 2 shows the overall architecture of our pipeline, which consists of two steps: (1) the NLU module which infers SMART attributes, dialogue acts, and phases; and (2) the goal extraction module which selects the SMART attribute values included in the patient's agreed-upon goal. In Figure 2, 'Goal-c' refers to the current goal; it starts with empty SMART attribute values and is updated as the week's messages are processed utterance-by-utterance. Below we will discuss the prediction models in the NLU module followed by the heuristics in the goal extraction module.[4]

4.1 Modeling SMART Attributes

This task involves predicting one of the 10 SMART attributes for each word or 'none'. We used Dataset 1 (27 patients) for modeling and performed 5-fold cross-validation (train/test: 22/5 patients). We experimented with both sequential and non-sequential classifiers such as CRF, Structured Perceptron (SP) (Collins, 2002), Logistic Regression (LR) (Grimm and Yarnold, 1995), Support Vector Machines (SVM) (Cortes and Vapnik, 1995), and Decision Trees (DT) (Quinlan, 1986). For features, we tried different combinations of - the current word, left and right context words, part-of-speech (POS) tags, left and right context words' POS tags, SpaCy named entity recognizer (NER), current word's phase, and ELMo word embeddings

[4]The phase and SMART attribute models are described in Gupta et al. (2020b) and briefly summarized here and in the appendix.

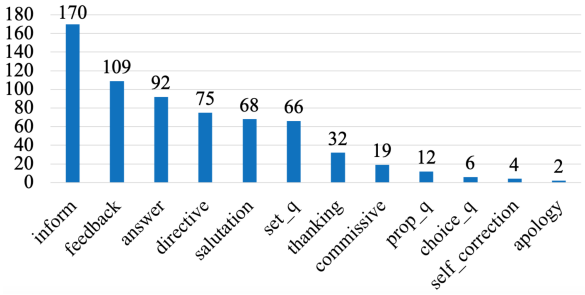

Figure 3: Dialogue acts distribution (15 weeks of data)

(Peters et al., 2018). The CRF, SP, and LR models performed the best without a significant difference between them using the current and context words, ELMo embeddings, and SpaCy NER. We decided to use the CRF model with an F1 macro score of 0.81. In our previous work (Gupta et al., 2020b), we used models with word2vec embeddings (Mikolov et al., 2013) but found ELMo embeddings to perform better.

4.2 Modeling Dialogue Acts

For DA prediction, we annotated 15 weeks (377 messages, 655 utterances) of goal setting data from Dataset 1 using the DAs schema described in the previous section.[5] The tag distribution is shown in Figure 3. Out of 655 annotated utterances, ≈89% of utterances are annotated with one of the 6 most

[5]We only annotated 15 weeks for DAs from six distinct patients due to the resource-intensive nature of human annotations. SMART attributes and phases annotations were done a couple of years earlier on the entire Dataset 1.

common DA tags. The remaining 11% consists of the other 6 DA tags. This shows the class/tag imbalance in the data. Though some of the tags are very rare, we still kept them as only a subset of data was annotated and they can be helpful for future annotations. We modeled DA prediction as a multi-class classification problem and experimented with CRF and five BERT-based models - two from Wu et al. (2020) and three from Cohan et al. (2019). BERT-base uncased model, a transformer self-attention encoder (Vaswani et al., 2017) with 12 layers and 12 attention heads with a hidden size of 768, was used for all the BERT-based models.

In Wu et al. (2020), the authors showed that task-oriented dialogue BERT (ToD-BERT), trained on nine human-human and multi-turn task-oriented datasets across over 60 domains, can perform better than BERT on tasks like DA classification, response selection, intent classification, and dialogue state tracking. Therefore, we use ToD-BERT for our dataset as well. In Cohan et al. (2019), the authors explored the use of BERT to jointly encode all the sentences in a sequence without the need for hierarchical encoding. The authors showed that jointly encoding the sentences for scientific abstract sentence classification task worked better than individual encoding followed by a transformer layer and CRF. Since context is important for dialogues, we decided to use their BERT sequential sentence classification (BERT SSC) model and hierarchical baseline models for our work as well.

- **CRF**: This model was given a sequence of utterances for one week as input and a sequence of DAs that maximizes the probability over the entire sequence was predicted. Features like BERT sentence embeddings, sender of the message, utterance length, distance of the message from the top in a week, presence/absence of a SMART attribute, previous utterance, and previous utterance embeddings were used in various combinations. The first four features together gave the best performance (F1 score macro = 0.68).
- **BERT**: Dialogue history was used as input, where a special [CLS] token was added in front of every input example, special tokens [SYS] and [USR] were appended in front of each coach and patient utterance respectively, and a [SEP] token was used between the history and the current utterance. E.g., [CLS] [SYS] S_1 [SYS] S_2 [USR] U_1 [SEP] [SYS] S_3, where S_i and U_i are the utterances from the messages. The DA tag for the current utterance S_3 was predicted using softmax function applied to [CLS] token encoding.
- **ToD-BERT**: Input and output representations were the same as for the BERT model above, except that the ToD-BERT masked language model was used for initialization.
- **BERT SSC**: One week of dialogue utterances were used as the model input. For the dialogues containing more than 10 utterances, the dialogue was recursively bisected until each split had less than or equal to 10 utterances (e.g., a dialogue with 27 utterances was divided into 3 groups of 9 utterances). Each utterance was separated by [SEP] token and a [CLS] was added in front of every input. The [SEP] token encodings were used to classify each utterance after it was passed through a multi-layer feedforward network.
- **BERT + Transformer layer (BERT-T)**: An utterance with a [CLS] token in front was passed as an input to BERT and the [CLS] token encoding was saved. These encoded representations were then collectively passed through an additional transformer layer to contextualize them over the entire sequence. After that, a final feedforward layer is used to generate a DA tag for each utterance. A maximum of 30 utterances was passed at a time through the transformer layer. If more, the data was divided recursively, like BERT SSC.
- **BERT + Transformer layer + CRF (BERT-T-CRF)**: In addition to the transformer layer, a CRF layer was also added after the feedforward layer above. The logits were passed through CRF to predict the DAs for the entire sequence. A maximum of 30 utterances was used here too.

For all the models above, 5-fold cross-validation was performed (train/test: 12/3 weeks). For all the five BERT based models, the test fold was used for early stopping. For training, we used a dropout ratio of 0.1, learning rate of $5e^{-5}$, Adam optimizer (Kingma and Ba, 2015), cross-entropy loss, batch size of 4, and 30 epochs. All the other parameters were kept the same as in the original papers (Cohan et al., 2019; Wu et al., 2020). We used the code publicly available for both papers on github.[6,7]

Google Colab free GPU (Tesla T4 ≈13GB RAM) was used for running the BERT-based models and CPU (2.6 GHz Dual-Core i5 8GB RAM)

[6] https://github.com/jasonwu0731/ToD-BERT

[7] https://github.com/allenai/sequential_sentence_classification

Model	all DAs	9 most frequent DAs	Runtime (mins)
BERT SSC	0.46	0.55	12
BERT-T	0.57	0.68	9
BERT-T-CRF	0.65*	0.76*	6
CRF	**0.68***	0.73*	4
BERT	0.66*	0.76*	28
ToD-BERT	**0.68***	**0.79***	22

Table 1: DA prediction F1 (macro) scores and average runtimes on Google Colab GPU, CRF on CPU

for the CRF model. The results for DA prediction are shown in Table 1. Statistical significance was calculated using ANOVA followed by posthoc Tukey tests (Tukey, 1949). A '*' in the table means that the corresponding model is significantly better than the BERT SSC model; the last four models, all better than BERT SSC, are statistically indistinguishable. The average train/test runtime over 5-folds was the lowest for the CRF model even with much slower hardware.

Our results contrast with the authors' observations in both papers (Cohan et al., 2019; Wu et al., 2020). First, both BERT and ToD-BERT performed almost the same, contrary to the original paper; this is possibly due to the difference between the health coaching dataset and the domains that ToD-BERT is trained on. Gururangan et al. (2020) showed the importance of domain adaptive pretraining as well. Second, the BERT-T-CRF model performed better than the BERT SSC model i.e. encoding individual utterances first and then contextualizing them performed better than passing all the utterances as input at the same time. The authors showed the opposite is true. However, their task was abstract sentence classification (non-dialogue data) and therefore, it is hard to compare the two. We might have observed a statistically significant difference with a larger dataset, but given the resource-intensive nature of manual annotations, we wanted to use minimally annotated data to show the applicability of these models. Of note is that BERT and ToD-BERT models will perform the same in an online setting as they only require dialogue history, but other models are set-up for an offline setting.

4.3 Modeling Phases

The task of phase prediction involved predicting one of the 6 phases for a given message. Since a phase like *refining* is more likely to follow *identification*, we explored both sequential and non-sequential classifiers such as CRF, SP, LR, SVM, and DT. Similar to SMART attributes, we used Dataset 1 (27 patients) and 5-fold cross-validation for modeling. We experimented with different combinations of features - unigrams, the distance of the message from the top, presence/absence of a SMART attribute, message length, normalized time difference between the current and previous message, the sender of the message, and word2vec word embeddings averaged over the entire message. CRF performed the best (F1 score macro = 0.71) using the first three features. We tried ELMo word embeddings as well, but embeddings as a feature did not help to improve the performance.

4.4 Extracting the Goal Summary

Next, we use the models described above for goal extraction. For phase and SMART attribute prediction, we used the CRF models and for DA prediction, we experimented with the four best performing models, but only present the results for the CRF and BERT models here. The phases model was retrained on the same 15 weeks that the DA model was trained on, for a fair comparison. We analyzed three different goal extraction methods.

1. **SMART** (baseline): We extracted the last mention for each of the 10 SMART attributes
2. **SMART+Phases**: We sequentially extracted SMART attributes from each message and updated the existing values unless the current message belonged to *follow-up* phase.
3. **SMART+DA**: We sequentially extracted SMART attributes from each utterance and updated the existing values unless the current utterance was an *inform* DA.

For evaluation, we used 30 goals/weeks (611 messages): 15 weeks from Dataset 1 (different from the ones annotated for DAs) and 15 weeks from Dataset 2 and compared the output against manually created gold standard goal summaries. For *activity* and *score* attributes, we took the last mention, as *activity* already had high accuracy and for *score*, we didn't notice an improvement. We also experimented with binary CRF classifiers for both phases (*follow-up* vs others) and DAs (*inform* vs others), but they did not improve performance for goal summarization. Additionally, binary classifiers would not be as useful for the dialogue agent.

Figure 4 shows the goal extraction performance for SMART attributes. We can observe that *amount* (e.g., 5000 steps), a crucial attribute, improves by 17.67% using the SMART+DA (BERT) model.

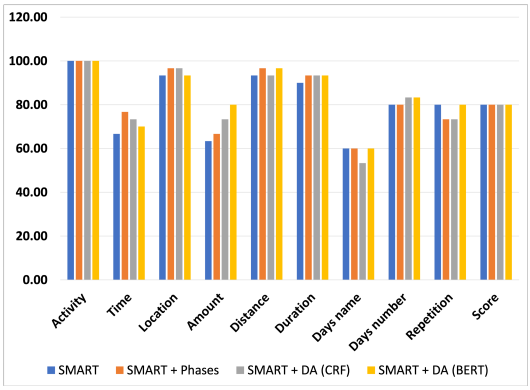

Figure 4: Percentage of SMART attributes correct

Both the SMART+DA models perform better than others for the *days number* attribute as well. For *distance* and *duration*, the two SMART+DA models and SMART+Phases model perform the same, but better than the SMART model. For *time* and *location*, SMART+Phases performed the best out of all the four models. Finally, for *repetition* and *days name*, both SMART and SMART+DA (BERT) performed the same. From these results, we can conclude that it is safe to use the SMART+DA (BERT) model for all the attributes as it always performed equal or better than the SMART model. When looking into SMART+Phases, we saw that it performed the best for two attributes, but also had a negative dip in performance for the *repetition* attribute. Therefore, we adopt the goal extraction pipeline that uses both dialogue acts (BERT) and phases as shown in Figure 2. Given the small performance difference on *time* and *location* between phases and DAs, to process messages in real-time, we will use only the SMART+DA (BERT) model, as it only requires the dialogue history. Additionally, to generate messages in real-time, the current *Goal-c* could be used. E.g., if *location* is null in *Goal-c*, the coach can ask for *location* next.

We previously showed in Gupta et al. (2020b) that metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are not appropriate for our extraction-based goal summaries as they are sensitive to exact word match (Reiter, 2018). That is, if a given word, say 'two', is classified as *days number* instead of *distance*, they will still output a high score as 'two' is in the reference summary. BLEU also favors shorter sentences, so missing attributes lead to a higher score.

	correct	partially correct	incorrect
C1	7	5	0
C2	2	8	2
C3	9	3	0

Table 2: Health coaches' evaluation of the summaries

5 Human Evaluation

Evaluating models with automatic metrics is important, but it is equally important to evaluate the usefulness and usability of these models with their users. We performed a pilot evaluation with the help of three health coaches to answer two main questions: (1) What is the health coaches' understanding of a correct goal summary? and (2) Are these goal summaries helpful?

We created an assessment using Google Forms and presented the three health coaches, who coached the patients during the Dataset 2 collection, the same 12 <set of messages-goal summary> pairs, where each pair consisted of a full set of weekly messages and the goal summary generated by our pipeline. The 12 pairs were chosen from 12 different patients, where each health coach had coached 4 of these patients. The summaries were generated by the SMART+Phases model as the evaluation took place before the DA prediction model was built. But we can expect the same if not better results in terms of coaches' feedback as the goal summaries have improved with DAs.

For each <set of messages-goal summary> pair, the coaches were asked to judge the given goal summary as correct, partially correct, or incorrect. In case of partially correct or incorrect, they were asked to write the correct goal summary. Partially correct meant some of the SMART attributes were missing a value whereas incorrect meant that some of the attributes had an incorrect value. The evaluation results are shown in Table 2. Coach 1 and coach 3 are similar in their evaluation, however, coach 2 found most goal summaries to be only partially correct. We found out that coach 2 was not clear on whether the goal of say '5000 steps Mon-Fri' meant 5000 steps each day or all together over the 5 days. Sometimes that information is not explicitly mentioned in the messages. The other two coaches assumed it to be for each day.

At the end of the assessment, the coaches were asked on a scale of 1 to 3, how useful a correct, partially correct, or incorrect goal summary would be to them. To this, all the three coaches said 3

(helpful) for correct goal summaries, 2 (neutral) for partially correct summaries, and 1 (not helpful) for incorrect summaries. This means that higher accuracy is required for the health coaches to feel comfortable in using goal summaries. The assessment form also consisted of an open-ended feedback field to write their overall impression of these goal summaries. One of the coaches said, "It would be nice to have the goal (summarized correctly) available and easily viewable, so that we would not have to scroll all the way backwards through our conversation and reread texts to figure out what the goal was. So thank you for doing this!".

6 Conclusions and Future Work

Many applications exist to promote a healthy lifestyle but they lack coaching components that are essential to keep the user motivated long term. In this paper, we discussed our work towards building a virtual assistant health coach that can help patients to set specific and realistic physical activity goals. Mainly, we focused on the goal summarization pipeline that is built upon the NLU module of the system and showed its usefulness for health coaches. We found that utterance-level information captured by dialogue acts improves goal summarization performance. Next, we will test its usability and helpfulness in an online setting while coaches are communicating with the patients in real-time. Following that, we will use phases, dialogue acts, and SMART attributes prediction models to generate possible responses for the coaches.

In this paper, we have presented an approach that takes advantage of traditional Machine Learning models, contemporary deep learning ones, and heuristics. We believe that for certain domains where accuracy of information is important, and data is scarce, such as the health coaching exchanges we have discussed here, end-to-end approaches are neither feasible, because of lack of large datasets, nor appropriate, since usability and usefulness for different types of stakeholders are crucial. We cannot claim that our mixed approach would work for any conversational agent in a health care or legal domain where scarce data is available; however, we would encourage researchers who work on such applications, to experiment with a variety of methods as we do here.

7 Acknowledgements

We would like to thank Nikolaos Agadakos (University of Illinois at Chicago) for insightful discussions. This work is supported by the National Science Foundation, initially by award IIS 1650900 and currently by award IIS 1838770.

References

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

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. *Dialogue acts in Verbmobil 2*. Citeseer.

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

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The HCRC map task corpus. *Language and speech*, 34(4):351–366.

Tatiana Anikina and Ivana Kruijff-Korbayova. 2019. Dialogue act classification in team communication for robot assisted disaster response. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 399–410, Stockholm, Sweden. Association for Computational Linguistics.

S Bauer, R Percevic, E Okon, R u Meermann, and H Kordy. 2003. Use of text messaging in the aftercare of patients with bulimia nervosa. *European Eating Disorders Review: The Professional Journal of the Eating Disorders Association*, 11(3):279–290.

Timothy W Bickmore, Everlyne Kimani, Ha Trinh, Alexandra Pusateri, Michael K Paasche-Orlow, and Jared W Magnani. 2018. Managing chronic conditions with a smartphone-based conversational virtual agent. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 119–124.

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

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.

Eric R Buhi, Tara E Trudnak, Mary P Martinasek, Alison B Oberne, Hollie J Fuhrmann, and Robert J McDermott. 2013. Mobile phone-based behavioural interventions for health: A systematic review. *Health Education Journal*, 72(5):564–583.

Harry Bunt. 2009. The dit++ taxonomy for functional dialogue markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, et al. 2010. Towards an iso standard for dialogue act annotation. In *Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Harry Bunt, Volha Petukhova, and Alex Chengyu Fang. 2017a. Revisiting the iso standard for dialogue act annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017b. Dialogue act annotation with the iso 24617-2 standard. In *Multimodal interaction with W3C standards*, pages 109–135. Springer.

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

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Daniel S Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3684–3690.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8.

Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56. Boston, MA.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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

Layla El Asri, Hannes Schulz, Shikhar Kr Sarma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR Mental Health*, 4(2).

Caroline Free, Gemma Phillips, Leandro Galli, Louise Watson, Lambert Felix, Phil Edwards, Vikram Patel, and Andy Haines. 2013. The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review. *PLoS Medicine*, 10(1).

Laurence G Grimm and Paul R Yarnold. 1995. *Reading and understanding multivariate statistics.* American Psychological Association.

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

Itika Gupta, Barbara Di Eugenio, Brian D Ziebart, Bing Liu, Ben S Gerber, and Lisa K Sharp. 2020b. Goal summarization for human-human health coaching dialogues. In *FLAIRS Conference*, pages 317–322.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Eun Young Ha, Joseph F Grafsgaard, Christopher M Mitchell, Kristy Elizabeth Boyer, and James C Lester. 2012. Combining verbal and nonverbal features to overcome the "information gap" in task-oriented dialogue. In *Proceedings of the 13th An-*

nual Meeting of the Special Interest Group on Discourse and Dialogue, pages 247–256. Association for Computational Linguistics.

Homa B Hashemi, Amir Asiaee, and Reiner Kraft. 2016. Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining, Workshop on Query Understanding*.

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.

Melinda H Huffman. 2009. Health coaching: a fresh, new approach to improve quality outcomes and compliance for patients with chronic conditions. *Home Healthcare Now*, 27(8):490–496.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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.

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

Paul Krebs, James O Prochaska, and Joseph S Rossi. 2010. A meta-analysis of computer-tailored interventions for health behavior change. *Preventive Medicine*, 51(3-4):214–221.

Abhinav Kumar, Barbara Di Eugenio, Jillian Aurisano, and Andrew Johnson. 2020. Augmenting small data to classify contextualized dialogue acts for exploratory visualization. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 590–599, Marseille, France. European Language Resources Association.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Geoffrey Leech and Martin Weisser. 2003. Generic speech act annotation for task-oriented dialogues. In *Proceedings of the Corpus Linguistics 2003 Conference*, volume 16, pages 441–446. Citeseer.

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

Edwin A Locke and Gary P Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9):705.

Audrey F Manley. 1996. *Physical activity and health: A report of the Surgeon General*. Diane Publishing.

Colleen M McBride and Barbara K Rimer. 1999. Using the telephone to improve health behavior and health service delivery. *Patient Education and Counseling*, 37(1):3–18.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 3111–3119.

WR Miller and S Rollnick. 2002. Motivational interviewing: preparing people for change. 2002. *New York: Guilford*, 2.

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

Stephen Palmer, Irene Tubbs, and Alison Whybrow. 2003. Health coaching to facilitate the promotion of healthy behaviour and achievement of health-related goals. *International Journal of Health Promotion and Education*, 41(3):91–93.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Shachi Paul, Rahul Goel, and Dilek Hakkani-Tür. 2019. Towards universal dialogue act tagging for task-oriented dialogues. *Proc. Interspeech 2019*, pages 1453–1457.

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 Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

Katrina L Piercy, Richard P Troiano, Rachel M Ballard, Susan A Carlson, Janet E Fulton, Deborah A Galuska, Stephanie M George, and Richard D Olson. 2018. The physical activity guidelines for americans. *JAMA*, 320(19):2020–2028.

JR Quinlan. 1986. Induction of decision trees. *Machine Learning*, 1(1):81–106.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Mickael Ringeval, Gerit Wagner, James Denford, Guy Paré, and Spyros Kitsiou. 2020. Fitbit-based interventions for healthy lifestyle outcomes: systematic review and meta-analysis. *Journal of Medical Internet Research*, 22(10).

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51.

Ameneh Shamekhi, Timothy Bickmore, Anna Lestoquoy, and Paula Gardiner. 2017. Augmenting group medical visits with conversational agents for stress management behavior change. In *International Conference on Persuasive Technology*, pages 55–67. Springer.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Mobile Fact Sheet. 2018. Pew research center. *URL: https://www.pewresearch.org/internet/fact-sheet/mobile/[accessed 2020-09-07]*.

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

Thomas Stephens. 1988. Physical activity and mental health in the united states and canada: evidence from four population surveys. *Preventive Medicine*, 17(1):35–47.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

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.

Pedro J Teixeira, Eliana V Carraça, David Markland, Marlene N Silva, and Richard M Ryan. 2012. Exercise, physical activity, and self-determination theory: a systematic review. *International Journal of Behavioral Nutrition and Physical Activity*, 9(1):78.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Chien-Sheng Wu, Steven CH Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289.

A Appendix

A.1 Details on the SMART Prediction Model

The performance for each SMART attribute is shown in Table 3. The SMART model uses the Conditional Random Fields (CRF) model with the feature combination of current word, the left and right context words, ELMo word embeddings, and SpaCy named entity recognizer.

Label	P	R	F1
Activity	0.952	0.956	0.952
Time	0.696	0.660	0.670
Location	0.787	0.757	0.747
Quantity-amount	0.946	0.922	0.934
Quantity-distance	0.700	0.554	0.594
Quantity-duration	0.886	0.950	0.906
Days-name	0.804	0.730	0.760
Days-number	0.834	0.820	0.816
Repetition	0.752	0.618	0.664
Attainability score	0.876	0.884	0.878
None	0.980	0.990	0.986
Macro average	0.838	0.804	0.810

Table 3: SMART attribute prediction results per label

A.2 Details on the Phase Prediction Model

Table 4 shows the results for each phase using the CRF model with the feature combination of unigrams, distance of the message from the top in a week, and SMART attributes.

Label	P	R	F1
Anticipate barrier	0.836	0.814	0.816
Follow up	0.908	0.922	0.912
Identification	0.816	0.858	0.828
Negotiation	0.482	0.360	0.368
Refining	0.660	0.732	0.678
Solve barrier	0.722	0.588	0.632
Macro average	0.738	0.712	0.708

Table 4: Phase prediction results per label

A.3 Details on Goal Extraction Results

Figure 5 shows the percentage of goals (y-axis) with given number of SMART attributes (x-axis) correctly extracted. Similar to the per attribute performance, the SMART+DA (BERT) model performed the best. It extracted 20% of goals (6 out of 30 goals) with all 10 attributes correct. On the other hand, the SMART+Phases and SMART (baseline) models only had 13.33% of goals (4 out of 30 goals) with all 10 attributes correct, and the SMART+DA (CRF) model only had 6.67% goals (2 out of 30 goals) correct. Going further down in the number of attributes, we found that both the CRF and BERT-based DA models had an equal percentage of goals (43.33%) with at least 9 attributes correct (adding percentages for 10 and 9 attributes correct). However, complete goal correctness is important for health coaches, therefore, the SMART+DA (BERT) model was chosen for the final goal extraction architecture.

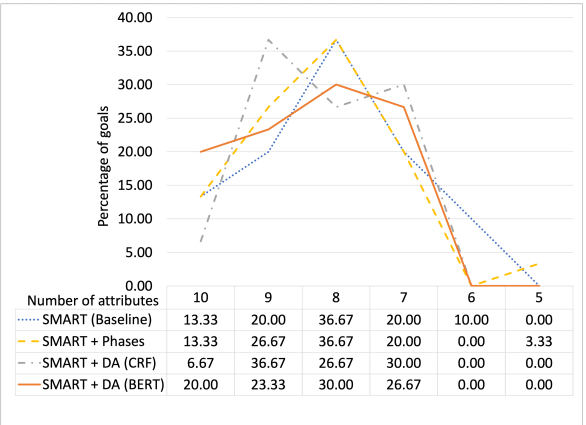

Figure 5: Percentage of goals with given number of attributes correct

A.4 Evaluation survey

Figure 6 shows an example from the evaluation survey given to the health coaches.

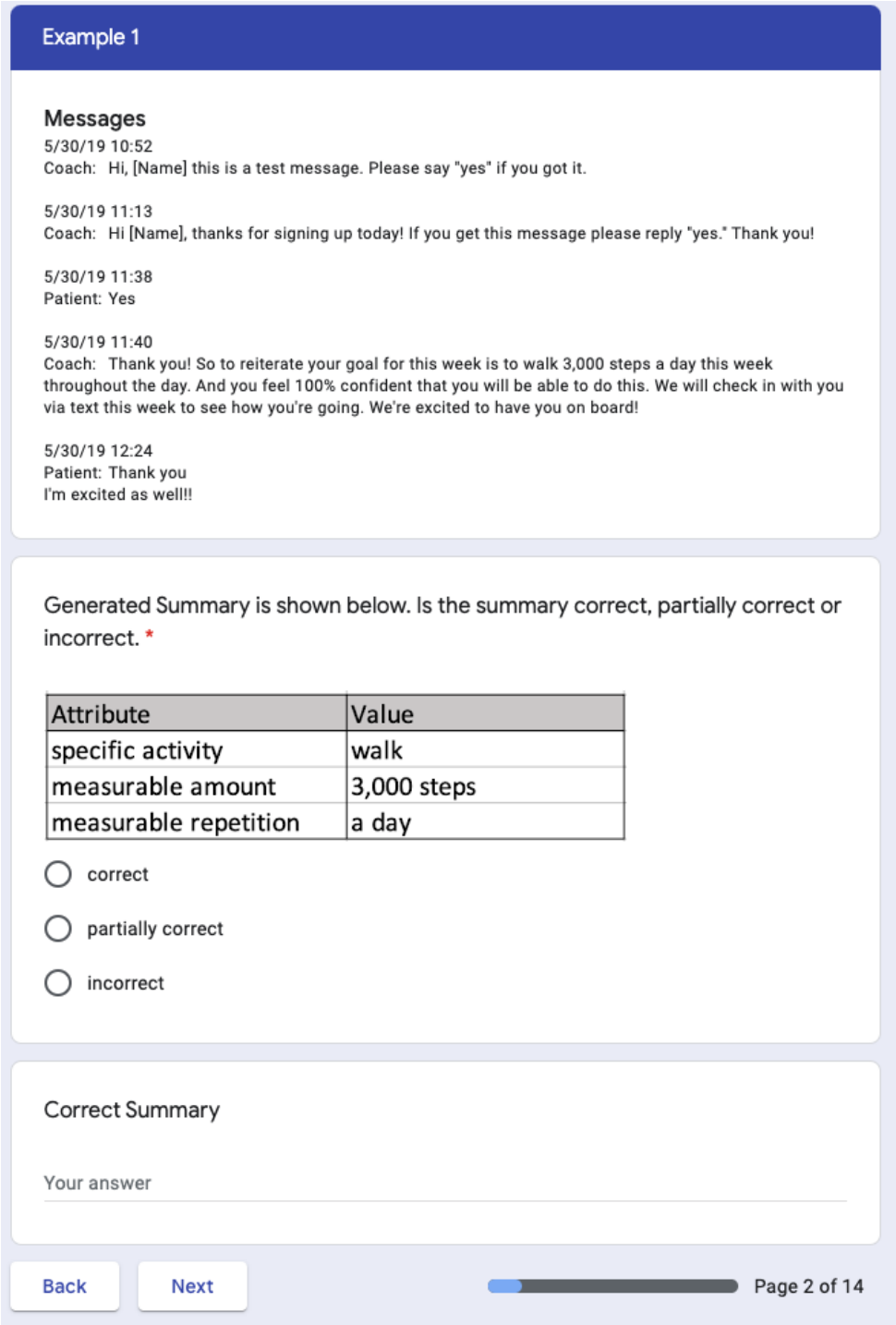

Figure 6: Example from the evaluation survey