Dear XYZ,

We addressed the feedbacks of the reviewers. In the original paper, the changed text is colored with red, if the change is not minor such as typo or grammatical problem. In this letter, we followed same approach with additional explanations which are bold.

This document organized as, the point of the reviewer and our solution for that. We received reviews of second reviewer in PDF format, while copying the content of it to word, some formatting issues occurred. However, it is very minimal.

Best Regards,

Ugan Yasavur and Christine Lisetti

================================================================================

Reviewer #1: The paper presents a RL-based dialog system for ECA to provide BI on drinking problems to increase awareness in the person and possibly identify people that may be at risk and need support. The dialog system will be integrated in a larger system embedded with extra capabilities to better understand the situation and lead the person to engage.  
  
Most of ECA models are rule driven and hence suffer from optimization. Applications are also tuned for short interactions. A reinforcement learning approach may provide a more effective system as it can better model the variety of the situation that may occur without having to explicitly identify them. It may also continuously improve as interactions occur. The system is fully evaluated on 89 participants showing the increased performances and acceptability of the system.  
The contribution is very important and goes beyond "drinking  problems".  In various health areas, prevention and self-directed rehabilitation are becoming main stream approaches due to the high cost and requirements imposed by the increasing number of people with health problems. An example is chronic pain condition. Increasing awareness of particular behaviour (e.g., anxiety due to wrong beliefs about the relation between pain and movement) may help reduce the risk of transitioning from acute to chronic pain e.g., during post-operation period. It may also help people with chronic pain to be more aware of how their own behaviour in response to pain. Physiotherapists with cognitive behavioural therapy engage in dialogs with their patients to increase awareness and teach skills and provide information to help people change their behaviour rather than correct the behaviour itself (Singh et al. 2014). Through bonding and dialog such capable avatars may better engage people in  
understanding their problems when clinical support is not available. The wider contribution of the system could be briefly discussed in the final discussion section. But this is just a suggestion.  
  
Singh, A, Klapper, A, Jia, J, Fidalgo, A, Tajadura-Jimenez, A, Kanakam, N, Bianchi-Berthouze, N, CdeC Williams, A (2014) Motivating People with Chronic Pain to do Physical Activity: Opportunities for Technology Design, CHI' 2014  
  
A few spelling mistakes and minor clarifications:  
-    It would be interesting to know why 18 was chosen as max number of questions. Is it the typical n questions in BI?

**We strictly followed NIAAA Brief intervention Guide [33] for adults. In the guide, there are total 18 questions to perform brief intervention. In section 3.1, we clarify the details of the NIAAA Brief Intervention guide.**

-    State the number of participants that used the testing system (I suppose 89-52)

**The information was exist in the text under section 4.2, we make the numbers bold. The reviewer’s supposition is correct. We also added the information to beginning of evaluation section (Section 4.1).**

In the second phase, the remaining 37 subject used the testing system.

-    Page 3, line 51: can usually handle less than 5, -->  can usually handle ARE less than 5,

**Fixed**  
-    Page 4, Line 33: Because of space limitations, we cannot not present a longer dialogue. --> Remove "not" …  I think.

**Fixed**  
-    Page 6: Line 27: the whole system into 5 sections according the BI guide --> the whole system into 5 sections according TO the BI guide

**Fixed**  
-    Page 6: Line 28: For the each step --> remove "THE"

**Fixed**  
-    Page 6, Line 35: In the each step -->  At each step (?)

**Fixed**  
-    Page 6, Line 50: ReAsk actions is -->  ReAsk actions are …. Please check the same for ASK actions and Configuration Actions.

**Fixed**  
-    Page 8, Line 27: that are updated by system -->  that are updated by THE system

**Fixed**  
-    Page 13 Line 23: the the -->  the

**We can’t find it.**

Reviwer : 2

Review on: 05/23/2014

Intelligent Virtual Agents and Spoken Dialog Systems come together to

Deliver Health Interventions

1. What is the purpose of this paper?

The authors present an “embodied conversational agent” that provides insight and aware- ness regrading alcohol problems by using “brief intervention” counseling approaches. The proposed system is based on:

(a) a 3D anthropomorphic speech-enabled interface – a spoken dialog interaction system between virtual agents and patients; and

(b) a spoken dialog system that is modeled based on Markov decision processes and the parameters of the models are learned using incremental off-line Q-Learning using data collected from real user interactions.

2. Is the method of approach valid?

Step 1 (asking): [5 questions]

Step 2 (assessing):

– Sub-step 2.1 (about abuse) : [4 questions][if 0 or 1 abuse indicator, then go to Sub-step 2.2]

– Sub-step 2.2 (about dependence): [7 questions][if 3 dependence indicators, then go to

Step 3 – Sub-step 3.2, otherwise, go to Step 3 – Sub-step 3.1]

Step 3 (advising):

– Sub-step 3.1 (at-risk): [1 question (reviewer inferred)][behavior modification]

– Sub-step 3.2 (disorder): [1 question (reviewer inferred)][treatment]

I inferred that there are two questions in Step 3. Is this correct? Otherwise, it is better to draw a diagram for subsection 3.1 and then explain it.

**The reviewer’s inference is correct. To clarify, we added additional clarifications about step 3 to in section 3.1.**

In Step 2, in the Assessment of Abuse stage, there are 4 questions assessing alcohol abuse indicators. It is enough to find one indicator of alcohol abuse to move to the assessment of dependence stage. If the system can not find any indicator of abuse with the 4 questions, it passes to the dependence stage. In the dependence stage, there are 7 questions. It is enough to detect 3 dependence indicators. If the system detects 3 dependence indicators, it transits to Step 3, drinkers with alcohol use disorder. Otherwise, it transits to the stage for at-risk drinkers. Therefore, the dialog branches to two separate steps in step 3; 1) one for at-risk drinkers, and **2**) one for drinkers with alcohol use disorder. In both branches, the system provides information related to the assessment of the system. If the system assesses that user has an alcohol use disorder, it refers the user to treatment and suggest a drinking goal after asking the intention of the user for changing. If the user is an at-risk drinker, it gauges his or her readiness to change, and provides feedback and information about the person's drinking. Therefore in both steps, the system provides factual information about person's drinking and suggested drinking limits, and asks user's intention to change with a single question. In total there can be a maximum of 18 different questions in a single session.

Subsection 3.2 : the term “dialog strategies” needs to be defined.

**We did by adding following sentence to section 3.2.**

In other words, a dialog strategy specifies, for each system state, what is the next action to be taken by the system.

The reference [25] needs to be a part of the related work. It should be clearly stated how the proposed method differs from reference [25], if there are any differences. Otherwise, it should be clearly mentioned that the methods are similar.

**[25] is very influential paper in SDS area, it is one of the earliest MDP-based dialog system. Because of its position, the ideas have influenced almost every MDP and POMDP based SDS. However, there is not direct connection to that paper. The reason of citing that paper in section 3.2, it explains and demonstrates very clearly dialog strategies in a very small dialog domain.**

**Even though there are many differences with [25], the most important one is in the way dialog strategy learning performed. They used user simulations from ATIS dialog corpus, in the beginning their system starts with no knowledge at all. In our system, even though it randomly selects actions as in [25], we constrained our system in each state to select 2 possible actions. In other words, our system knows in the beginning selecting which actions make sense in a particular state. It enabled us to achieve optimization goals faster. It might not be possible to learn approximately optimal dialog strategies, if we followed the same way. The reason is selecting randomly one of the available 169 dialog actions will create extremely large number (169680 , 680 is the number of the states) of random dialog strategies and learning the good strategies will take a lot time. In the case of [25], in ATIS task with total 7 dialog actions, it takes 710,000 dialogs to learn optimal dialog strategy. However, we are not questioning approach in [25], it is done in that way to assess possibility of automatic dialog strategy learning from scratch. In our case, we already know that it is possible, we applied it to real-world problem. We also need to be more practical due to the problem of not having data.**

**We added following text to section 2.2. under related research.**

Unlike the very classic dialog strategy learning approaches [25] in which the system literally has no knowledge for dialog action selection in the training stage, our system knows taking which actions make sense in each state despite being non-optimal as in [44]. For example, taking a farewell action in the beginning of dialog instead of greeting does not make sense. Our approach enables our system to learn dialog strategies faster from small amount of dialog corpus than the systems that has absolutely no knowledge in the training.

The definition of T is erroneous. It should be T(s, a, s0) = P(s0|s, a). The sentence “... which describes how the probability of performing action a in state s0 will lead to state s

...” must be fixed as “... which describes how the probability of performing action a in state s will lead to state s0 ...”.

Fixed the error, w used P for probability notation.

R(s, s0) should be replaced with “R(s, a)” or “R(s, a, s0)” such that R(s, a) =

P

a P(a|s)R(s, a, s0)

Fixed.

In the context of RL, MDP is also defined by γ ∈ [0, 1], the discount factor. Therefore, it is recommended to include γ in the given definition.

We described γ after the formula 1.

There are 18 questions. Each question corresponds to 34 states: therefore, there are 612 + 1 states, if you include the lumped terminal states. For each stage, there are 170 + 136 + 238 + 68 + 68 = 680 states, which is = 18 × 34 = 612. The author say that there are 578 states, which is confusing!

**We made a mistake in that paragraph. We re-calculated and re-wrote the related paragraph.**

The number of states are 170, 136, 238, 68 and 68 respectively for step 1 ask, step 2 abuse, step 2 dependence, step 3 for at-risk drinkers, and step 3 for drinkers with alcohol use disorders. Total number of states is 680.

18 × 9 = 162. I don’t understand the sentence: “18 questions multiplied by the number of available actions, except no-confirmation action”. This sentence needs to be rephrased or modified. The total number of actions are 162 + 2 × 2 + 3 = 169.

**We make that paragraph clearer by adding new sentences and re-organizing.**

The available actions for the first question in \textit{Step 1} (Asking About Alcohol Use) is shown in Table \ref{dialogActions}. The number of available actions for each question is 9 as for the first question. Although the length of the dialogue is not fixed, our system asks a maximum of 18 questions. There are 162 available actions (for asking and re-asking questions, and confirmations) for system to select in longest dialog session (18 questions multiplied by the number of available actions). There are dialogue actions which are used while transiting from one step to another step (e.g. from Step 1 to Step 2 Abuse) and dialog actions for ending the conversation.

The sentence, “In other words, the remaining 4 actions are nonterminal states in which the system needs to wait for a response from user before it can proceed ”, is ambiguous. Action states are not nonterminal states. What is the semantic behind this sentence?

**We addressed the confusion. We added additional sentences and removed the word nonterminal.**

As we mentioned before, for each question there are 34 states. State updates are performed based on user's dialog actions or systems dialog actions in each dialog turn. In Table 4, only 30 state-actions mappings are shown that are updated by the system dialog actions or user dialog actions. The remaining 4 states are only updated based on user dialog actions, that's why we did not include them in Table 4. The reason for this is that, if the system waits for the confirmation from user (i.e. in other words C=5 see Table 2), the system dialog actions can not be used to update a state. In other words, the remaining 4 states needs to be updated by user dialog actions. In Table 4, we only showed the states that are updated by system.

The sentence, “Our system can learn approximately optimal dialog strategies for the ini- tiative style and the confirmation type selection”, can only be stated after the experiments are conducted. As the approach does not provide an analysis of this kind, the only way to confirm the above statement is to conduct the experiments.

**We re-write the sentence.**

Our system aims to learn approximately optimal dialog strategies for the initiative style and the confirmation type selection.

Section 3.4: technically, it is not “To avoid the data sparsity problem ...”. It is more appropriate to state “To avoid the curse of dimensionality ...”.

Fixed

Equation (1) should be:

Q∗(s, a) = R(s, a) + γ X P(s0 s, a) max Q∗(s0, a0)

|

a0

s0

and ~~where~~ ~~T(s,~~ ~~a)~~ ~~is~~ ~~the~~ ~~transition~~ ~~model~~ ~~and~~ ~~R(s,~~ ~~a)~~ ~~is~~ ~~the~~ ~~local~~ ~~reward~~ ~~function.~~.

Fixed

It is not clear from the text, whether the algorithm executes in learning or planning mode. In the learning mode, you do not need to know the transition probabilities or expected reward received. In the planning mode, on the other hand, you have to have these values, or need to learn the model first, before executing the algorithm. A clarification is necessary.

**We call learning mode as training mode, and planning mode as testing mode. In training mode we do not use the algorithm, it is just for data collection, in the testing mode the algorithm is used to approximately optimal dialog policies. We added the text below.**

As we describe in Section 4.1, we run the systems in two modes, training/exploration and testing. Training mode is for data collection, in testing mode the system uses optimized dialog strategies based on the data collected in training. Therefore, Equation 1 is used only for testing mode.

If the dynamics of the conversations could be modeled using ODEs with the data at hand, then the batch mode reinforcement learning techniques have reported to produce effective results (e.g., [ESGW06], [GVAP08]). This could be a potential future extension to the work. The given reward functions are acceptable.

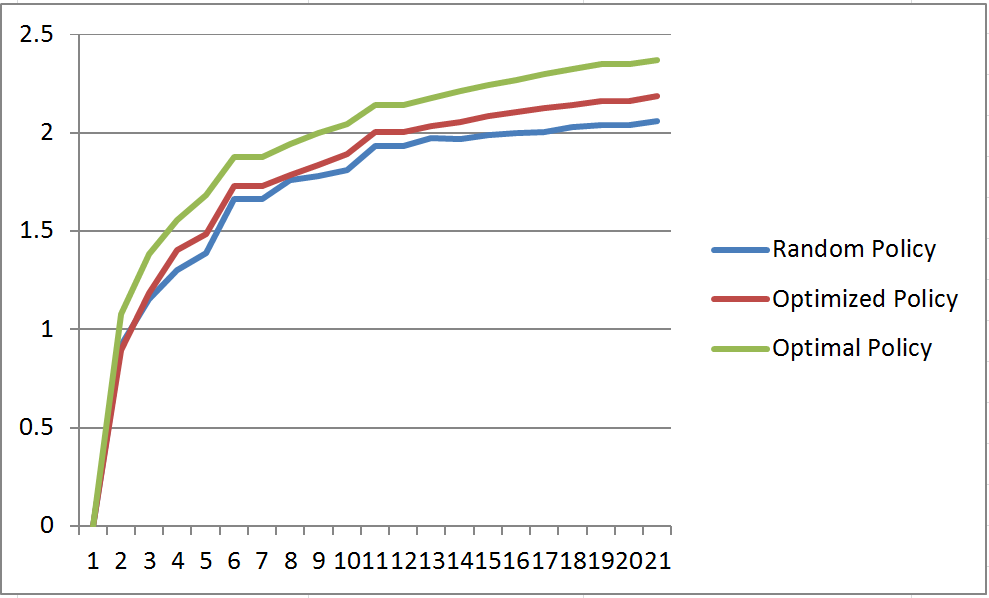
**We will take into account recommendations of the reviewer for the future extensions.**

3. Is the actual execution of the research correct? Are the correct conclusion being drawn from the results?

Since the main focus of the paper is to show that the optimal strategies can indeed be learned from tabular based Q-Learning, the authors need to show a few graphs for a ques- tion that actually showing the improvement of the Q-values. If the learning curve asymp- totes for the given data, we know that the system has learned optimal values (because this paper uses tabular representation – if the project were to use function approximation, this statement no longer holds, and Q-Learning in fact be divergent. In such a situation, one should consider on-line off-policy algorithms such as GQ(λ) [MS10]). This would give guidance in re-playing and accumulating more data. The baseline can be a hand-designed solution (or may be random).

**We create a diagram which shows that Q-values in each episode/question. Since our time steps are based on progress of the intervention, we depicted log-scale graph which demonstrates Q-values. You can see the added text and diagram below.**

**However, we compared Q-values for each episode. An episode can be defined as completing one question and passing to the next question. Completion of a question does not mean that obtaining information that the system tries to get. It is possible to transit to next question without obtaining the information, in that case system receive negative reward. We described in Section 3.6 details of reward function. We depicted improvement of Q-values for each episode in Fig. 4. The reason of having 21 episodes is 18 questions plus transitions between MDPs. As it is shown in Fig. 4, the optimized policy performed better, even though it is not optimal. We have to note that *optimal policy* shows highest reward that the system can achieve, whereas *random policy* and *optimized policy* is the average score that the system collected in training and testing operation modes respectively.**

****

Since the paper states that the RL method has found the optimal solution, it is very important to show that indeed the optimal solution has been found. Otherwise, it is better to say that the “best solution” is found. It is not clear from the evaluation that optimal solution is found. But, it shows that statistically significant results are found. Therefore, it is important to distinguish the differences.

**Actually we did not claim that we find optimal solutions solution but we find approximately optimal solutions. With the new information and diagram, we clarified this point better.**

4. Is the presentation satisfactory?

The paper is well written. There are a few typos and notational errors present in the paper as listed below:

1. The quotation marks on LaTeX follow the standard “foo bar”. The paper uses ”foo bar”. It is better to use LaTeX standard as it is widely recognized.

**Fixed**

(b) Replace “dialogue” with “dialog” in all places. If the paper needs to be in British

English, then swap the order.

**Fixed**

(c) Is “embodied conversational agent” a synonym for “intelligent virtual character”?

Page 2, subsection 2.1, last paragraph: replace “(or embodied conversational agent ECA)” with “(or embodied conversational agent (ECA))”. ECA abbreviation has been introduced in page 1. It is not necessary to abbreviate it again.

**Fixed**

(d) Page 2, subsection 2.2, fist paragraph: replace “(based on reinforcement learning)” with “(e.g., based on RL)”. Technically, RL is an iterative/batch stochastic approxi- mation to the expected return. Therefore, the term “statistical” is ambiguous. There- fore, I would replace “... which can be statistical (based on reinforcement learning), or hand-crafted. ...” with “ ... which could be either based on an area of machine learning (e.g., based on RL) or hand-crafted. ...”.

**Fixed**

(e) Technically, there is no such thing called “statistical reinforcement learning (RL)”.

Replace “Systems based on statistical reinforcement learning (RL) ...” with “Systems based on RL ...”.

**Fixed**

(f ) Section 3.3, paragraph 2: “... For each state, there are 5 common features, Question, Confidence, Value and Grammar and Aux ...” should be “... For each state, there are 5 common features, Question, Confidence, Value, Grammar, and Aux ...”.

**Fixed**

(g) Section 3.3, paragraph 2: “... step-specific requirement in ~~the~~ each step ...”.

**Fixed**

(h) Section 3.3, paragraph 3: ASR is not defined.

**Fixed**

(i) Section 3.3, paragraph 3: A comma after “ ... the type of grammar is restrictive

(Gr=0)”

**Fixed**

(j) Section 3.3, paragraph 3: Polarity of alcohol problem indicator is not defined.

**Fixed**

(k) Section 3.3, paragraph 6: “Our system uses 2 types of initiative dialog actions, system initiative and user initiative ...”. Replace the “,” with “:”.

**Fixed**

(l) Section 3.3: “The reason for this is that, if the system waits for the confirmation from user(i.e. ... ”. Added space before “(”.

**Fixed**

(m) Section 3.4: Is it “... between the system and a client ...” or “... between the system and a patient ...”?

**Fixed, this was in section 3.3., we make it “between the system and a user”**

(n) Section 3.4: Instead of “... multiple goal states ...”, it is better to say “... multiple terminal states ...”. Some terminal states terminate the Step (such as the consent state), and some terminal states provides transparent transitions to another start state (or start state distribution) of another MDP.

**Fixed, section 3.4 last paragraph.**

(o) Section 3.5: The sentence, “At the beginning, we did not have any data as dialog corpus because none exist in the domain of alcohol abuse.”, must be rephrased. One potential solution could be “At the inception of the project, we did not have data or state transition triples for our domain of discourse (domain of alcohol abuse).”

**Fixed**

At the inception of the project, we did not have any data for optimizing the system for our domain of discourse (domain of alcohol use).

5. Overall how good is it? What do you recommend?

The paper is an innovative contribution to the given domain of discourse (domain of alcohol abuse), and it has many potential usages and practical value. There are few typos and technical errors present in the paper as stated above. I recommend to accept the paper for publication once the above mentioned issues are resolved.

References

[ESGW06] Damien Ernst, G-B Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In Decision and Control, 2006 45th IEEE Conference on, pages 667–672. IEEE, 2006.

[GVAP08] Arthur Guez, Robert D Vincent, Massimo Avoli, and Joelle Pineau. Adaptive treat- ment of epilepsy via batch-mode reinforcement learning. In AAAI, pages 1671–1678,

2008.

[MS10] H. R. Maei and R. S. Sutton. GQ(λ): A general gradient algorithm for temporal- difference prediction learning with eligibility traces. In Conference on Artificial General Intelligence, 2010.