A Wizard-of-Oz Study on A Non-Task-Oriented Dialog Systems That Reacts to User Engagement

Zhou Yu, Leah Nicolich-Henkin, Alan W Black and Alex I. Rudnicky
School of Computer Science
Carnegie Mellon University
{zhouyu,leah.nh,awb,air}@cs.cmu.edu

Abstract

In this paper, we describe a system that reacts to both possible system breakdowns and low user engagement with a set of conversational strategies. These general strategies reduce the number of inappropriate responses and produce better user engagement. We also found that a system that reacts to both possible system breakdowns and low user engagement is rated by both experts and non-experts as having better overall user engagement compared to a system that only reacts to possible system breakdowns. We argue that for non-task-oriented systems we should optimize on both system response appropriateness and user engagement. We also found that apart from making the system response appropriate, funny and provocative responses can also lead to better user engagement. On the other hand, short appropriate responses, such as "Yes" or "No" can lead to decreased user engagement. We will use these findings to further improve our system.

1 Introduction

Non-task-oriented conversational systems do not have a stated goal to work towards. Nevertheless, they are useful for many purposes, such as keeping elderly people company and helping second language learners improve conversation and communication skills. More importantly, they can be combined with task-oriented systems to act as a transition smoother or a rapport builder for complex tasks that require user cooperation. They have potential wide use in education, medical and service domains.

There are a variety of existing methods to generate responses for non-task-oriented systems, such as machine translation (Ritter et al., 2011), retrieval-based response selection (Banchs and Li, 2012), and sequence-to-sequence recurrent neural network (Vinyals and Le, 2015). All aim to improve system coherence, but none of them focus on the experience of the user. Conversation is an interaction that involves two parties, so only improving the system side of the conversation is insufficient. In an extreme case, if the system is always appropriate, but is a boring and passive conversational partner, users would not stay interested in the conversation or come back a second time. Thus we argue that user engagement should be considered a critical part of a functional system. Previous researchers found that users who completed a task with a system but disliked the experience would not come back to use the system a second time. In a non-task-oriented system, the user experience is even more crucial, because the ultimate goal is to keep users in the interaction as long as possible, or have them come back as frequently as possible. Previously systems have not tried to improve user experience, mostly because these systems are text-based, and do not have access to the user's behaviors aside from typed text. In this paper, we define user engagement as the interest to continue the conversation in each turn. We study the construct using a multimodal dialog system that is able to process and produce audio-visual behaviors. Making the system aware of user engagement is considered crucial in creating user stickiness in interaction designs. Better user engagement leads to a better experience, and in turn attracts repeat users. We argue that a good system should not only be coherent and appropriate but should also be engaging.

We describe a multimodal non-task-oriented conversational system that optimizes its performance on both system coherence and user engagement. The system reacts to both user engagement and system generation confidence in real time us-

ing a set of active conversational strategies. System generation confidence is defined as the confidence that the generated response is considered appropriate with respect to the previous user utterance. Although the user engagement metric is produced by an expert in a Wizard-of-Oz setting, it is the first step towards a fully automated engagement reactive system. Previously very little research addressed reactive systems due to the difficulty of modeling the users and the lack of audiovisual data. We also make the audiovisual data along with the annotations available.

2 Related Work

Many experiments have shown that an agent reacting to a user's behavior or internal state leads to better user experience. In an in-car navigation setting, a system that reacts to the user's cognitive load was shown to have better user experience (Kousidis et al., 2014). In a tutoring setting, a system that reacts to the user's disengagement resulted in better learning gain (Forbes-Riley and Litman, 2012). In task-oriented systems users have a concrete reason to interact with the system. However, in a non-task-oriented setting, user engagement is the sole reason for the user to stay in the conversation, making it an ideal situation for engagement study. In this paper, we focus on making the system reactive to user engagement in real time in an everyday chatting setting.

In human-human conversations, engagement has been studied extensively. Engagement is considered important in designing interactive systems. Some believe engagement is correlated with immersiveness (Lombard and Ditton, 1997). For example, how immersed a user is in the interaction plays a key role in measuring the interaction quality. Some believe engagement is related to the level of psychological presence (i.e. focus) during a certain task (Abadi et al., 2013), for example how long the user is focused on the robot (Moshkina et al., 2014). Some define engagement as "the value a participant in an interaction attributes to the goal of being together with the other participant(s) and of continuing the interaction" (Peters et al., 2005). In this paper, we define engagement as the interest to continue the conversation. Because the goal of a non-task-oriented system is to keep the user interacting with the system voluntarily, making users have the interest to continue is critical.

A lot of conversational strategies have been proposed in previous work to avoid generating incoherent utterances in non-task-oriented conversations, such as introducing topics, (e.g. "Let's talk about favorite foods!" in (Higashinaka et al., 2014)) and asking the user to explain missing words. (Schmidt et al., 2015). In this paper, we propose a set of strategies that actively deal with both user engagement and system response appropriateness.

3 System Design and User Experiment Setting

The base system used is Multimodal TickTock, which generates system responses by retrieving the most similar utterance in a conversation database using a key word matching method (Yu et al., 2015). It takes spoken utterances from the user as input and produces synthesized speech as output. A cartoon face signals whether it is speaking or not, and can present some basic expressions. This clearly artificial design aims to avoid the uncanny valley dilemma, so that the users do not expect realistic human-like behaviors from the system. It has the capability to collect and extract audio-visual features, such as head and face movement (Baltrusaitis et al., 2012), in real time. These features are not used in this experiment, but will be incorporated as part of automatic engagement recognition in the future.

We designed five strategies based on previous literature to deal with possible system breakdowns and to improve user engagement.

1. **Switch Topics** (switch): propose a new topic other than the current topic, such as "Let's talk about sports."

2. **Initiate activities** (initiation): propose an activity to do together, such as "Do you want to see the latest Star Wars movie together?".

3. **End topics with an open question** (end): close the current topic using an open question, such as "Could you tell me something interesting?".

4. **Tell A Joke** (joke): tell a joke such as: "Politicians and diapers have one thing in common. They should both be changed regularly, and for the same reason.".

5. **Refer Back to A Previously Engaged Topic** (refer back): refer back to the previous engaging topic. We keep a list of utterances that have resulted in high user engagement. This strategy will refer the user back to the most recently engaged turn. For example: "Previously, you said 'I like music', do you want to talk more about that?"

Each strategy has a set of surface forms to choose from in order to avoid repetition. For example, the *switch* strategy has several forms, such as, "How about we talk about sports?" and "Let's talk about sports."

We designed two versions of Multimodal TickTock: REL and REL+ENG. The REL system uses the strategies above to deal with low system generation confidence (system breakdown). The generation confidence is the weighted score of matching key words between the user input and the chosen utterance from the database. The REL+ENG system uses the strategies to deal with low system generation confidence, and in addition reacts to low user engagement. One caveat is that the *refer back* strategy is not available for the REL system. In the REL+ENG system, a trained expert annotates the user's engagement as soon as the user finishes the utterance. A randomly selected strategy triggers whenever the user engagement is 'Strongly Disengaged' or 'Disengaged'. Any non-task-oriented system can adopt the above policy and strategies with minor system adjustments.

For systems that use other response generation methods, the confidence score can be computed using other metrics. For example, a neural network generation system (Vinyals and Le, 2015) can use the posterior probability for the confidence score.

In order to avoid culture and language proficiency confound, all participants in the study are originally from North America. Gender was balanced as well. We had 10 people (6 males) interact with REL and 12 people (7 males) interact with REL+ENG. Participants were all university students and none of them had interacted with a multimodal dialog system before. There are no repeat users in the two groups. We also collected how frequently they use spoken dialog systems, such as Apple Siri, in the after-experiment user survey in the REL+ENG study, and found that 25% of them have used dialog systems frequently. In the future, we hope to collect a more balanced dataset to test this factor's influence.

An example dialog of a user interacting with REL is showfn in Table 1. In Turn 5, the system used the *switch* strategy to deal with the fact that it was unable to retrieve a response with similar content to the user's utterance. An example dialog from REL+ENG is shown in Table 2. In Turn 3, the system used the *initiation* strategy to deal with the fact that the user was not engaged in the previous turn. Figure 1 is a snapshot of an example video recording of one of the interactions. The blue box represents the head-tracking results, and the red dots represent face-tracking. Together with the audio-visual recordings, we have also made all the annotations and manual transcriptions available [1].

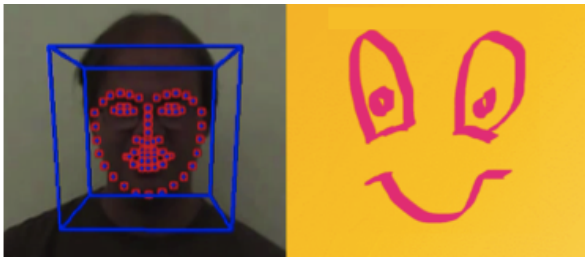

Figure 1: Screen shot of an example conversation recording

4 Annotations

We annotated each conversational exchange with user engagement and system response appropriateness. One conversational exchange includes one speech segment from the system and the following speech segment from the user. The engagement is a rating based on the user's performance during the entire turn, while the system's appropriateness is a rating with respect to the user's utterance in the previous turn. User engagement is defined as "user's interest to continue the conversation", and the system appropriateness as "the coherence and appropriateness of the system utterance given the context". Table 3 and 4 describe the user engagement annotation scheme and the system appropriateness annotation scheme respectively. For turns when users failed to respond, we label the engagement as 0. We labeled the turns that the system failed to understand the user and asked the user to repeat to be 0 in appropriateness. We exclude all these turns in the analysis

[1] www.cmuticktock.org

Turn	Participant	TickTock
1	Your noise is a question mark, I think that's interesting.	Of course, what's your idea?
2	I think it fits you perfectly.	Let me see, his nose is still broke.
3	Do you want me to make your nose broken?	Ah, yes, why did you ask me for this?
4	Because I thought you might mind.	Well, I am not getting into specifics about what doctor treated him, how he was treated.
5	You have not been treated yet. That is in the future.	Let me see, by the way, are you interested in movies?
6	Yes, I watch movies sometimes.	Great.

Table 1: An example conversation in REL

Turn	Participant	TickTock
1	How are you	I think so, any thoughts?
2	How are you doing?	I am fine, how are you, any thought?
3	Any thought?	Let me see, do you want to go to my friends' party tonight?
4	No I have a lot of work to do. I can't go.	Everyday?

Table 2: An example conversation in REL+ENG

in order to avoid adding system failure in the analysis. The inter-annotator agreement between the two experts is high (kappa = 0.73) in appropriateness, but the agreement between the two experts in annotating engagement is relatively lower (kappa =0.43). However, if we collapse the rating for "strongly disengaged" with "disengaged" and "very engaged" with "engaged", our agreement improves (kappa = 0.67).

5 Evaluation

To evaluate the efficacy of the strategies, we also conducted an Amazon Mechanical Turk study to test if a non-expert would agree with the experts about which system elicits better user engagement. We selected video recordings with participants who are not familiar with dialog systems. There are only five participants in the REL dataset and nine participants in the REL+ENG dataset who meet this requirement. In order to balance the two sets, we randomly selected five participants from the nine in the REL+ENG. We picked one video from each dataset to form a A/B comparison study. In total there are 25 pairs, and we recruited three raters for each pair. Nobody rated the same pair twice. We ask them to watch the two videos and then compare them through a set of questions including "Which system resulted in a better user experience?", "Which system would you rather interact with?" and "Which person seemed more enthusiastic about talking to the system". In addition, we also included some factual question related to the video content in order to test if the rater had watched the video, which all of them had. Raters are allowed to watch the two videos multiple times. The limitations of such a comparison is that some system failures, such as ASR failure, may affect the quality of the conversation, which may be a confound. In the task, we specifically asked the users to overlook these system defects, but they still commented on these issues in their feedback. We will collect more examples in the future to balance the influence of system defects.

6 Quantitative Analysis and Results

In this section, we first discuss whether the designed strategies are useful in avoiding system inappropriateness and improving user engagement. Then, we discuss whether both experts and non-experts who watched the video recordings of the interactions prefer a system that reacts to both low user engagement and system inappropriateness over a system that only reacts to low system appropriateness. In addition, we discuss the relationship between system appropriateness and user engagement. In the end, we discuss the relationship and methods to elicit user engagement and user experience.

6.1 Strategies and System Appropriateness

We found that designed conversational strategies are useful in avoiding system breakdowns. The system randomly selects one of the strategies described in Section 4 whenever its confidence in generating an appropriate answer is extremely low. In Table 5, we show for both REL and REL+ENG, how many times each strategy is triggered to react to low confidence in generating system responses and the distribution of the produced utterances being rated as "Inappropriate", "Interpretable" and "Appropriate". Among them, 63% and 73% of the turns are rated as "Interpretable" or "Appropriate"

Label	Definition	Description
1	Strongly Disengaged	Shows no interest in the conversation, not responding or engaged in other things.
2	Disengaged	Shows little interest to continue the conversation, passively interacts with his conversational partner.
3	Neither Disengaged nor Engaged	Interacts with the conversational partner, showing neither interest nor lack of interest to continue the conversation.
4	Engaged	Shows mild interest to continue the conversation.
5	Strongly Engaged	Shows a lot of interest to continue the conversation and actively contributes to the conversation.

Table 3: Engagement annotation scale and definition.

Label	Definition	Example
Inappropriate (1)	Not coherent with the user utterance	*Participant*: How old are you? *TickTock*: Apple.
Interpretable (2)	Related and can be interpreted	*Participant*: How old are you? *TickTock*: That's too big a question for me to answer.
Appropriate (3)	Coherent with the user utterance	*Participant*: How is the weather today? *TickTock*: Very good.

Table 4: Appropriateness rating scheme.

in REL and REL+ENG respectively. The percentage is higher in REL+ENG than REL mostly due to the introduction of *refer back* strategy, which the REL system could not use because it does not track the user's engagement. Compared to REL, which doesn't react to low system response generation confidence, REL+ENG successfully made 69% of inappropriate turns to be "Interpretable" or "Appropriate".

Each strategy has a different effect on improving the system's appropriateness. Among them, the *refer back* strategy leads to more appropriate responses in general, but happens infrequently, due to its strict trigger condition. It can only be triggered if the user previously had a high engagement utterance. The *initiation* strategy leads to more interpretable responses overall, because utterances like "Do you want to go to my friend's party?" actively seek user consent. Even though it may seem abrupt in some contexts, the transition will usually be considered to be interpretable. The *joke* strategy has a high probability of being inappropriate. However, if the joke fits the context, it may be appropriate. For example,

TickTock: "Let's talk about politics."

User: "I don't know too much about politics."

TickTock: "Let me tell you something, politicians and diapers have one thing in common, they both need to be changed regularly."

However, if the joke is out of the context, it will leave the participant with an impression that TickTock is saying random things.

In the future, we intend to track the topic of the conversation, and design specific jokes with respect to conversation topic. We intend to design additional strategies, such as performing grounding requests on out-of-vocabulary words (Schmidt et al., 2015), to address possible system breakdowns, and we will also implement a policy to control when to use which strategy.

6.2 Strategies and User Engagement

We found that designed conversational strategies are useful in improving user engagement. We created an engagement change metric that measures the difference between the current turn engagement and the previous turn engagement. In Table 6, we list the user engagement change for when each strategy triggered in the REL+ENG dataset. In total, 72% of the time when the system reacts to low user engagement, it leads to positive engagement change. We believe this is because the strategies we designed have an active tone, which can reduce the cognitive load required to actively come up with something to say. In addition, since these strategies are triggered when the user engagement is low, the random chance of them improving user engagement is already high, so the percentage of improving user engagement is even

	REL				REL+ENG			
Strategy	Total	InApp	Inter	App	Total	InApp	Inter	App
switch	46	13(28%)	27(59%)	6(13%)	32	6(19%)	18(56%)	8(25%)
initiation	10	2(20%)	6(60%)	2(20%)	18	0(0%)	8(44%)	10(56%)
end	29	14(48%)	13(45%)	2(17%)	16	6(38%)	8(50%)	2(13%)
joke	10	5(50%)	2(20%)	3(30%)	20	14(70%)	0(0%)	6(30%)
refer back	-	-	-	-	12	0(0%)	6(50%)	6(50%)
Total	95	34(35%)	48(51%)	13(14%)	98	26(27%)	40(41%)	32(33%)

Table 5: System appropriateness distribution when two systems react to possible system breakdowns.

higher.

For each strategy, the chance of improving the user's engagement is different. The *refer back* strategy is the most effective strategy: 75% of the time, it leads to better user engagement. We believe this is because once the system refers back to what the user said before, the user feels that the agent is somewhat intelligent and in turn increases his/her interest to continue the conversation, to find what else the system can do. For the *switch* and *end* strategies, there are examples of them both reducing and increasing user engagement. When we looked at the specific cases where the user engagement decreased, we found that those utterances are rated as inappropriate given the context. This leads us to believe that during the selection of what strategies we should use to react to user's low engagement, we should also consider whether the system utterance would be appropriate. We also examined the turns that did not improve or decreased user engagement and found that they are towards the end of the conversation, when the user lost interest and ended the conversation regardless of what the system said.

Strategy	Total	$\Delta < 0$	$\Delta = 0$	$\Delta > 0$
switch	10	1(10%)	3(30%)	6(60%)
initiation	5	0(0%)	2(40%)	3(60%)
end	3	1(33%)	1(33%)	1(33%)
joke	4	0(0%)	2(50%)	2(50%)
refer back	4	0(0%)	1(25%)	3(75%)
Total	26	2 (6%)	9(22%)	15(72%)

Table 6: User engagement change distribution when system reacts to low user engagement.

6.3 Third-person Preference

In our study, we found that a system that reacts to low user engagement and possible system breakdowns is rated as having better user engagement and experience compared to a system that only reacts to possible system breakdowns. This rating held true for both experts and non-experts. We performed an unbalanced Student's t-test on expert-rated user engagement of turns in REL and REL+ENG and found the engagement ratings are statistically different ($p < 0.05$). REL+ENG has more user engagement (REL: Mean = 3.09 (SD = 0.62); REL+ENG: Mean = 3.51 (SD = 0.78). A t-test on utterances that are not produced by designed strategies shows the two systems are not statistically different in terms of user engagement ($p = 0.13$). This suggests that the difference in user engagement is mostly due to the utterances that are produced by strategies. Experts also rated the interaction for overall user experience and we found that REL+ENG interactions are rated significantly higher than REL system overall ($p < 0.05$).

In REL+ENG, 37% of the strategies were triggered to react to low user engagement and 63% were used to deal with low generation confidence. Among the strategies that were triggered to react to low user engagement, 72% of them lead to user engagement improvement. We believe the ability to react to low user engagement is the reason that REL+ENG has more user engagement than REL. Another reason is that REL+ENG has an extra strategy, *refer back*, which in general performs best in improving user engagement. In the user survey, one of the participants also mentioned that he likes the REL+ENG system because it actively proposes engaging topics.

For non-expert ratings, there are 25 A/B comparison tasks. Each task had three raters, and we used the majority vote of the three raters as the final result. People rate REL+ENG as more engaging in 12 tasks, and REL more engaging in 3 tasks. Ten tasks were rated the same for both systems. For non-experts who watched the videos of the interactions, the REL+ENG system elicited

significantly more user engagement than the REL system. This conclusion is also true when the systems are judged on which leads to a better user experience. We examined the three tasks on which the REL system is rated higher than REL+ENG and found that two of them involved the same interaction produced by REL. In that interaction, the user is very actively interpreting the system's utterance and responding with interesting questions. Table 1 shows a part of that interaction.

6.4 System Appropriateness and User Engagement

In the conversations produced using REL, an unbalanced Student's t-test of engagement change between turns that are appropriate and ones that are inappropriate shows that turns that are appropriate (Mean =0.33, (SD=0.84)) have significantly ($p = -0.01$) better engagement change than turns that are inappropriate (Mean = -0.01, (SD=0.92)). Figure 2 shows a box plot of the resulting engagement change from appropriate and inappropriate responses. The figure suggests that having appropriate responses leads to better engagement change overall. However some inappropriate responses lead to positive engagement change as well. The same trend is found in conversations produced by REL+ENG.

We tested the hypothesis with respect to each strategy via an unbalanced Student's t-test. The hypothesis holds for the *switch*, *initiation* and *joke* strategies. It did not hold for the *end* strategy, but this is probably because there were very few examples of *end* being triggered and rated appropriate, making it hard to yield any statistical significance. In addition, across all responses, we find some outliers, where even though the system's response is appropriate the user's engagement decreased. This may happen when the system provides a simple 'yes' or 'no' answer, when the system interrupts the user, or when the user misunderstands the system. Some users are not familiar with synthetic voices and misheard the system, and thus thought the system was inappropriate.

We believe that in the future we can improve our system's turn-taking mechanism and try to tune the system retrieval method to prefer longer responses. This will help to overcome the issue that even appropriate answers can lead to a decrease in user engagement. Since appropriate system responses make users more engaged, are all the positive engagement changes the result of appropriate responses? We performed an unbalanced t-test of the appropriateness values between turns that have positive engagement change (Mean = 1.79 (SD = 0.82)) and turns that have negative engagement change (Mean = 1.53 (SD = 0.67)) and found that they are statistically significant ($p < 0.05$). We examined the recordings of conversations and found that there are other factors that contribute to the engagement change other than the system's appropriateness. For example, funny comments and provocative utterances on the part of the system can also increase user engagement. In Table 1, the system response in Turn 4 is only rated as "Interpretable," and yet it leads to an increase in user engagement. The speaker even smiled when replying to the system. In another interaction, "Don't talk to an idiot, because they will drag you down to the same level and beat you with experience." is rated as "Inappropriate" with respect to the previous user utterance. However the user reacted to it with increased engagement and asked the system: "Are you calling me an idiot, TickTock?". We conclude that being appropriate is important to achieve better user engagement, however it is not the only way.

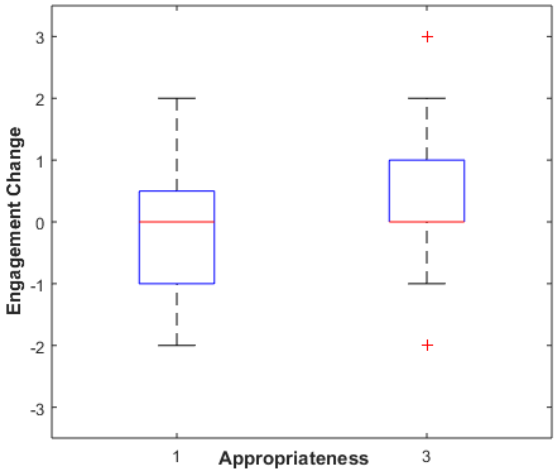

Figure 2: User engagement change with respect to system appropriateness in REL.

6.5 User Engagement and User Experience

In the survey after the REL+ENG study, we asked three questions to test the relationships among users' overall interaction engagement, users' positivity towards the agent, and users' overall experience in interacting with the system. We used a five-point Likert scale (1-5). The higher the score

is, the more engaged the user is, and the more positive the user is towards the system, the better the user experience the user has. We designed the survey carefully so these three questions are not next to each another, in order to avoid people's tendency to equate these questions. Exact matches between users' rating on their overall engagement (Mean = 2.75 (SD = 0.75)) and their positivity towards the system are found. This is surprising yet possible, since normal users may not differentiate between the two questions: "How engaged you felt during the interaction?" and "How positive you felt towards TickTock during the interaction?". They may internalize that being positive to your partner is the same as being engaged in the conversation. The overall user experience (Mean = 2.83 (SD = 0.71)) is also highly correlated (ρ = 0.92) with both user engagement and user positivity towards the system. Our finding suggests that improving user engagement is critical to eliciting better user experience in an everyday chatting setting. However, our sample size (12) is relatively small, and we plan to include more users in the study in the future.

Another question is whether users really know what "user experience" is. In future studies, we plan to include questions that are more specific such as, "Would you want to interact with the system again?", "Would you invite your friend to interact with the system?" and "Do you think the system is easy to talk to?".

7 Qualitative Results

After the users interacted with REL+ENG, we asked them to fill out a survey. We asked the users what they liked and disliked about the system, and for their suggestions for how to improve the system. A number of participants commented on the visual aspects of the system, mentioning that they liked the cartoon face and that it smiles a lot. Two participants said they liked the system because it actively proposes engaging topics, and tells jokes. This supports our hypothesis that our designed strategies are useful in increasing user engagement. Three users disliked the system because of its incoherence. Two users could not understand the synthesizer very well, which made them unsure whether answers were inappropriate or whether they had simply misunderstood the system. Two participants also complained that the system interrupted them sometimes and one participant mentioned that the system changes topics too often.

One participant suggested displaying subtitles below TickTock's face so that people would be able to comprehend the system's utterances better. Another participant proposed that TickTock should start the conversation with a topic to discuss in order to avoid the cognitive load imposed by the user's coming up with topics. We will consider both suggestions in our future studies.

8 Conclusion and Future Work

We designed and deployed a non-task-oriented conversational system powered with a set of designed strategies that not only reacts to possible system breakdowns but also monitors user engagement in real time in a Wizard-of-Oz implementation. The system reacts to user engagement or system breakdown respectively by randomly selecting one of the designed strategies whenever the user's engagement is low or the system's response generation confidence is low. In the study, our designed strategies are shown to be useful in increasing the system's appropriateness as well as in increasing the user's engagement. We found that appropriateness leads to better user engagement. However not all improved user engagement is elicited by appropriate responses. Sometimes, provocative and funny responses also work.

In a third-person study, experts rated the system that reacts to both low user engagement and low generation confidence as having more overall user engagement than the system that only reacts to low generation confidence. In an Amazon Mechanical Turk study, we found non-experts agreed with experts. We conclude that the improvement gained by reacting to user's engagement is generally recognizable.

One caveat is that due to the lack of a user survey in the REL study, we could not directly compare the self-reported engagement or user experience to determine which system is better. Thus, we plan to ask people to interact with both systems and report which system they like better and which system they think is more engaging.

We will implement an automatic engagement predictor in the real-time system to replace the human in the loop. In addition, a better policy to select strategies based on both user engagement and system response appropriateness will be developed.

References

Mojtaba Khomami Abadi, Jacopo Staiano, Alessandro Cappelletti, Massimo Zancanaro, and Nicu Sebe. 2013. Multimodal engagement classification for affective cinema. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013, Geneva, Switzerland, September 2-5, 2013*, pages 411–416.

Tadas Baltrusaitis, Peter Robinson, and L Morency. 2012. 3D constrained local model for rigid and non-rigid facial tracking. In *Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2610–2617. IEEE.

Rafael E Banchs and Haizhou Li. 2012. Iris: a chat-oriented dialogue system based on the vector space model. In *Proceedings of the ACL 2012 System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Katherine Forbes-Riley and Diane J. Litman. 2012. Adapting to multiple affective states in spoken dialogue. In *Proceedings of the SIGDIAL, Seoul National University, Seoul, South Korea*, pages 217–226.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *COLING*, pages 928–939.

Spyros Kousidis, Casey Kennington, Timo Baumann, Hendrik Buschmeier, Stefan Kopp, and David Schlangen. 2014. A multimodal in-car dialogue system that tracks the driver's attention. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 26–33. ACM.

M Lombard and T Ditton. 1997. At the heart of it all: The concept of presence, journal of computer mediated-communication. *Journal of Computer Mediated Communication*, 3(2).

Lilia Moshkina, Susan Trickett, and J. Gregory Trafton. 2014. Social engagement in public places: A tale of one robot. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, HRI '14, pages 382–389, New York, NY, USA. ACM.

Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. 2005. A model of attention and interest using gaze behavior. In *Intelligent Virtual Agents, 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005, Proceedings*, pages 229–240.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Maria Schmidt, Jan Niehues, and Alex Waibel. 2015. Towards an open-domain social dialog system. In *Proceedings of the 6th International Workshop Series on Spoken Dialog Systems*, pages 124–129.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. ICML Deep Learning Workshop 2015.

Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. 2015. TickTock: A non-goal-oriented multimodal dialog system with engagement awareness. In *Proceedings of the AAAI Spring Symposium*.