TickTock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness

Zhou Yu, Alexandros Papangelis, Alexander Rudnicky
Carnegie Mellon University, School of Computer Science
5000 Forbes Avenue, Pittsburgh, PA 15213
{zhouyu, apapa, air}@cs.cmu.edu

Abstract

We describe TickTock, a conversational agent designed to engage humans on topics of its choosing and to carry on an interaction for as long as possible. Our prototype uses a database of talk show transcripts featuring guests from the film industry. To be an interesting companion Tick Tock uses immediate context from the last two turns to formulate queries into a database of utterances. The process is automatic. TickTock monitors user engagement and performs certain moves, such as topic shifts, based on its assessment of user state. Initially we used utterance content for monitoring. We have subsequently begun to investigate non-language cues, such as prosody and visual cues to create a more robust engagement model based on multiple human communication channels.

Introduction

Human-human communication is highly reactive, with participants expressing goals, developing common ground and monitoring attention while maintaining an implicit social relationship. Automated agents, on the other hand, are designed to focus on the task, which could lead to dull and less engaging interactions with a human. Even worse, people might get frustrated with the linearity of the conversation and lose interest. This is an even greater problem when the interaction is machine-initiated and the human has no intention to interact in the first place. For example, consider a robot that wanders in a public space and tries to ask people questions. Intuitively, one would expect the robot to know how to engage the human in a conversation and keep it going by using conversation strategies commonly used by humans.

A pre-requisite of achieving that goal is to have some sense of your interlocutors' state of engagement. Are they in a rush? Are they paying attention? Did I just say something that's not right? A good conversation partner will track this state and rapidly modify their strategies to preserve engagement to allow the conversation to continue.

To study engagement, we developed a system that is capable of conducting free-form conversations, in contrast to goal-driven systems, which are designed to acquire information, provide feedback, or negotiate constraints with the human. A free-conversation system in principle removes any built-in value for the human and its success depends on the machine keeping the human interested in the ongoing conversation. Thus, as task completion is no longer an applicable metric, we chose to focus on the length of the conversation in time and the number of conversation turns exchanged. These metrics provide an objective criterion for engagement: what is the likelihood of the human contributing one more turn at any point of time in the conversation? This provides a basis for treating the problem in a machine learning framework.

Related Work

From Eliza (Weizenbaum, 1966), a hand-crafted rule based chatterbot system to Microsoft XiaoBing, a widely used social media chatterbot, a great deal of effort has been put into automatic dialog generation for non-goal oriented agents. However, little has been done in considering the user's mental state, such as engagement in dialog generation for a non-goal oriented agent. Corrigan et al. (2014) developed a robot that is aware of both the task and social engagement in a specific education task. Bohus and Horvitz (2009) developed a virtual human which is able to complete a specific task and is able to handle multiple human users.

Engagement or involvement of participants in human-human face-to-face conversations has been studied extensively. Gatica-Perez et al. (2005) asked judges to annotate interest level and group involvement over 15 second intervals in a four party dialogue on a 5 point scale. They define group involvement as "*the perceived degree of interest or* *involvement of the majority of the group”* and used HMMs over both speech and visual features of a multimodal corpus to detect segments of high and neutral group interest level. Bednarik and Hradis (2012) investigated 6 different levels of engagement (no interest, following, responding, conversing, influencing, managing) annotated by at least two annotators for 15 second intervals, and their relation to different gaze patterns.

Oertel et al. (2011) used an 11-point scale and also a time fixed annotation unit, which is 5 seconds long and then binned the annotation into three classes for modeling. They found that acoustic features, such as, high voice level, span and voice intensity and visual features such as, larger body movement are predictive of involvement. Mutual gaze is more predictive of involvement compared to audio features (Oertel et al., 2012). Bonin et al. (2012) relied on annotators’ intuition to come up with a definition of involvement instead of giving a formal definition. They proposed to not predefine chunks in which annotators are supposed to give ratings but rather ask the annotators to identify the point at which involvement changes. Levitski et al. (2012) carried out engagement annotations for the silent participant in a three party conversation to study nonverbal behaviors. A binary distinction was made between “engaged” and “passive”. The state “engaged” was annotated when the person actively gazes at the other participants while “passive” was annotated when the person’s gaze and gesture indicate less involved in the conversation. They found that participant gaze focused more on the background when the silent interlocutor is perceived as “passive”.

Significant effort was put into annotating engagement, both in definition and annotation units, also in linking human signals, both acoustic and visual, to engagement, and in predicting engagement using those signals in human-human conversations.

Our work focuses on developing a foundation for sensing and using engagement in human-machine conversation. Non-goal oriented conversation provides the best paradigm for this investigation as we believe it attenuates the role of factors that come up in goal-oriented dialog. Once understood, engagement management can be incorporated into goal-oriented dialog systems.

System Overview

Figure 1 shows the architecture of the TickTock conversational agent; we use Google Automatic Speech Recognition (ASR), MultiSense and purpose-build Natural Language Understanding (NLU) to process the input, a RavenClaw-based dialogue manager and template-based Natural Language Generation (NLG), Flite Text to Speech (TTS) and an animated head for realization. An earlier version of a free-form conversational agent guided some of our design decisions (Marge et al., 2010).

Question – Answer Database

Our database consists of question-answer pairs from CNN Interview Transcripts from the “Piers Morgan Tonight” Show[1]. The corpus has 767 Interviews in total and each interview is between 500 to 1,000 sentences. To construct our database, we used a rule-based question identification method, which simply means searching for tokens such as “?”, “How”, “Wh-”, etc. to identify questions and then extracted the consecutive utterance of the other speaker as the answer to that question.

Answer Retrieval

User speech is decoded using Google ASR. Human acoustic and visual signals are captured by MultiSense (Scherer, 2012) and are used to estimate engagement as well as detect user presence. The ASR result is processed by the NLU component, we first do POS tagging (Toutanova, 2003) and remove stop words; heuristics are then used to compute weights, e.g. nouns have higher weight compared to other POS tags. We then search for each keyword in the database and calculate the weighted sum, which becomes the retrieval confidence score. Finally, we normalize the score by diving it by the length of the retrieved utterance. We filter out inappropriate content, excluding the retrieved answer if it is longer than 15 words and remove other characters such as parentheses or square brackets (along with everything between them).

Key Term Matching (Martin, 2002) was used for content retrieval. Our goal is to generate coherent responses efficiently without deep understanding of the context, which is useful in a non-task oriented interactive system, and is motivated by lexical cohesion in modeling discourse. The coherence can be reflected by the repetition of lexicon items. The method first does shallow syntactic analysis of the input utterance and extracts keywords. These are used to search the corpus for a suitable response.

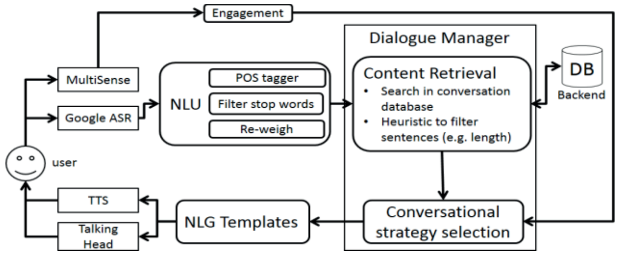

Figure 1: The architecture of TickTock

[1] http://transcripts.cnn.com/TRANSCRIPTS/

Conversational Strategies

Once we retrieve content, we select a conversational strategy, based on a heuristic (implemented in the Engagement module), i.e. a pre-defined threshold for the retrieval confidence score, which can be tuned to make the system appear more active or more passive. Higher thresholds correspond to more active systems. We designed two strategies for each case and at each dialogue turn, we randomly chose between the two strategies.

If the retrieval confidence score is low, meaning no good response was obtained, we use strategies to change the current topic by proposing a new topic, such as "sports" or "music" or we close the current topic using an open question, such as "Could you tell me something interesting?" If the retrieval confidence score is higher than the threshold, we choose between two strategies, one where we return the retrieved answer and another where we ask about the user's opinion, for example: "What do you think?" These simple strategies are designed as baselines to sustain the flow of conversation.

Content Realization

The content is realized via NLG templates that can embody conversational strategies, e.g. "What do you think of [music]?" Flite TTS (Black, 2005) is used for synthesis. TickTock uses a 2D talking head composed of 6 animated line drawings Pappu et al. (2013). These are used to explicitly communicate internal state. Initially, when there is no user interacting with the system, the talking head is in a semi-active state and appears to be dozing. Once it hears the attention phrase ("listen to me"), it wakes up; the lips move when the agent is speaking and it looks confused when non- or mis-understanding occurs.

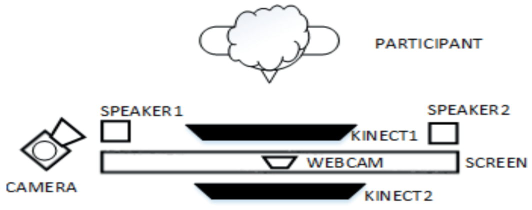

Figure 2: Experiment Setting

Data Collection

The experiment involved participants, freely interacting with TickTock over the course of 5 minutes, one at a time. The setting is shown in Figure 2. The participant stood in front of the system, with the talking head displayed on a screen suspended at the participant's eye-level. A separate video recording was also made, and used as the prompt for the engagement annotation task. The system collected different channels of speech data, including the segmented utterances, the corresponding Google ASR output, the TTS output, as well as timing information. For visual data, we used MultiSense, to collect face feature points, face orientation and location, skeleton joint location, as well as some high level features, such as body lean (forward, backward). This is an example of a dialog:

TickTock: What do you think of movies?
Human: What do I think of movies? I like movies.
TickTock: Well, can you share some interesting stuff with me?
Human: Sure...Let's see…I...[long pause] am not sure.
TickTock: let me see, by the way, are you interested in sports?

Engagement Annotation

As defined by Sidner et al. (2004), engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. In this study, we use a five-point engagement annotation scale. The main component we measure is how much one would like to continue interacting with the system (see Table 1 for details). The unit of the annotation is the conversational turn, defined as the start of the system's utterance to just before the next system utterance. Unlike most previous work, instead of using a fixed length of time, we chose conversational turn, which captures the micro-dynamics of discourse and is also the minimum planning point for the dialog manager.

Level	Definition	Description
1	Strongly disengaged	Shows no interest in dialogue system, engaged in other things than talking to the dialogue system
2	Disengaged	Shows little interest to continue the conversation, passively interacts with the dialogue system
3	Neither disengaged nor engaged	Interacts with the dialogue system, 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 1 Engagement annotation scale and definition

Engagement, by its very nature, is difficult to annotate, as it reflects an internal state of the participant. To establish a baseline standard we asked participants, immediately after their session, to watch the video recording and mark their level of engagement for each conversational turn. The scale used is shown in Table 1 and is based on Sidner et al. (2004). To assess whether another person could reliably judge engagement, we had a second judge (one of the authors) also annotate the recording. Our goal was two-fold: (1) to obtain a "gold standard" annotation that is as accurate as possible,and (2) to determine whether an observer could reliably annotate the state of engagement. Success with our second goal would be positive evidence for the plausibility of training models to classify engagement levels.

Analysis

We asked a third person to annotate the interaction and compared the two versions. Due to iterative modification of the annotation manual and of the experiment setup, we report only the three interactions collected according to the finalized setting and annotation scheme. There are 149 utterances, 25 utterances were excluded due to annotator errors. Figure 2 shows the distribution of engagement scores over the 124 utterances with valid annotations in both self-reported and third person versions. The average turn duration is 7.75 seconds.

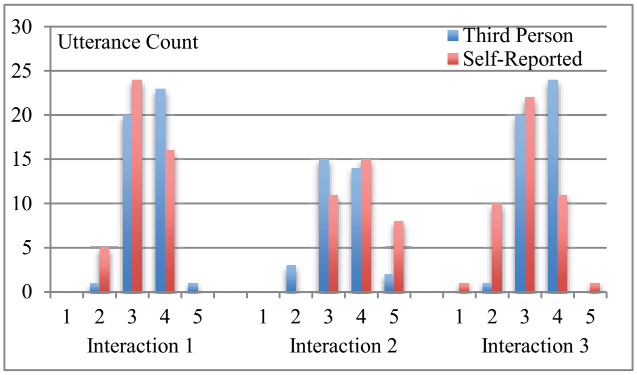

Figure 3: The Distribution of Engagement score for self-reported and third person annotation.

For 13 utterances (10.5%) with valid annotations, the self-reported the third-person annotations differed more than a single level. Specifically, we found that in the first and third interactions, the third person rates engagement consistently higher than the self-reported level, while in the second interaction, the third person annotation is consistently lower than self-reported level. Consulting the video record we found that users appear more facially expressive in interactions with higher third-person ratings.

Future Work

TickTock provides a framework for investigating non-goal oriented dialog systems with real time multimodal sensing. We plan to continue collecting data, possibly further modifying the collection and annotation schemes. One goal is to resolve the question of whether first-party or third-party annotation is more reliable. We will test reliability in modeling engagement using the behavioral data that we are collecting. We expect to model user engagement in real time using speech and image features. This will allow us to investigate how to condition agent interaction strategies to maintain user engagement level.

Acknowledgements

We would like to thank Yipei Wang, Yajie Miao, Kenneth Huang, Aasish Pappu and Ran Zhao for their contributions to earlier versions of the TickTock.

References

Bednarik R. and Hradis. M. 2005. *Gaze and conversational engagement in multiparty video conversation:An annotation scheme and classification of high and low levels of engagement*. Workshop on Eye Gaze in Intelligent Human Machine Interaction.

Black, A.W. and Lenzo, K.A. 2001. *"Flite: a small fast run-time synthesis engine."* In 4th ISCA ITRW on Speech Synthesis.

Bohus, D., Horvitz, E., 2009. *Models for Multiparty Engagement in Open-World Dialog*, SIGdial'09, London, UK

Bonin, F., Bock R. and Campbell, N. 2012. *How do we react to context? Annotation of individual and group engagement in a video corpus*. In International Conference on Social Computing.

Gatica-Perez, D., McCowan, I., Zhang, D. and Bengio, S. 2005. *Detecting group interest-level in meetings.* In Proc. of ICASSP,

Levitski, A., Radun, J. and Jokinen, K. 2012. *Visual interaction and conversational activity.* In Workshop on Eye Gaze in Intelligent Human Machine Interaction

McCowan, I., Gatica-Perez, D., Bengio, S., Lathoud, G., Barnard, M. and Zhang, D. 2005. *Automatic analysis of multimodal group actions in meetings.* Pattern Analysis and Machine Intelligence

Pappu, A., Sun, M., Sridharan, S. and Rudnicky, A. I. 2013. *Situated Multiparty Interactions between Humans and Agents.* Proceedings of HCII, July 2013, Las Vegas, NV.

Lee J. Corrigan, Christina Basedow, Dennis Küster, Arvid Kappas, Christopher Peters, and Ginevra Castellano. 2014. *Mixing implicit and explicit probes: finding a ground truth for engagement in social human-robot interactions*. HRI '14. New York, NY, USA

Marge, M., Miranda, J. Black, A.W. and Rudnicky, A. I. 2010. *Towards Improving the Naturalness of Social Conversations with Dialogue Systems*. Proceedings of SIGdial, Tokyo, Japan.

Martin, J. R. 2002. *Meaning beyond the clause: area: self-perspectives*. Annual Review of Applied Linguistics 22.

Oertel, C., Scherer, S and Campbell, N. 2012. *On the use of multimodal cues for the prediction of degrees of involvement in spontaneous conversation.* In Interspeech, Florence, Italy.

Scherer, S., Marsella, S., Stratou, G., Xu, Y., Morbini, Egan, F.A., Rizzo, A., and Morency, L.P. 2012. *Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors*, In Proceedings of Intelligent Virtual Agents

Sidner, C.L., Kidd, C.D., Lee, D., and Lesh, N. 2004. *Where to Look: A Study of Human-Robot Engagement*, In Proceedings of IUI, Madeira, Portugal.

Toutanova, K., Klein, D., Manning, C., and Singer, Y. 2003. *Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network.* In Proceedings of HLT-NAACL, pp. 252-259.

Weizenbaum, J., 1966, *"ELIZA—A Computer Program For the Study of Natural Language Communication between Man and Machine"*, Communications of the ACM 9 (1): 36–45