

A Demonstration of Dialogue Processing in SimSensei Kiosk

Fabrizio Morbini, David DeVault, Kallirroi Georgila,
Ron Artstein, David Traum, Louis-Philippe Morency

USC Institute for Creative Technologies

12015 Waterfront Dr., Playa Vista, CA 90094

{morbini, devault, kgeorgila, artstein, traum, morency}@ict.usc.edu

Abstract

This demonstration highlights the dialogue processing in SimSensei Kiosk, a virtual human dialogue system that conducts interviews related to psychological distress conditions such as depression, anxiety, and post-traumatic stress disorder (PTSD). The dialogue processing in SimSensei Kiosk allows the system to conduct coherent spoken interviews of human users that are 15-25 minutes in length, and in which users feel comfortable talking and openly sharing information. We present the design of the individual dialogue components, and show examples of natural conversation flow between the system and users, including expressions of empathy, follow-up responses and continuation prompts, and turn-taking.

1 Introduction

This demonstration highlights the dialogue processing in SimSensei Kiosk, a virtual human dialogue system that conducts interviews related to psychological distress conditions such as depression, anxiety, and post-traumatic stress disorder (PTSD) (DeVault et al., 2014). SimSensei Kiosk has two main functions – a virtual human called Ellie (pictured in Figure 1), who converses with a user in a spoken, semi-structured interview, and a multimodal perception system which analyzes the user’s behavior in real time to identify indicators of psychological distress.

The system has been designed and developed over two years using a series of face-to-face, Wizard-of-Oz, and automated system studies involving more than 350 human participants (Scherer et al., 2013; DeVault et al., 2013; DeVault et al., 2014). Agent design has been guided by two overarching goals: (1) the agent should make



Figure 1: Ellie, the virtual human interviewer in SimSensei Kiosk.

the user feel comfortable talking and openly sharing information, and at the same time (2) the agent should create interactional situations that support the automatic assessment of verbal and nonverbal behaviors correlated with psychological distress. During an interview, the agent presents a set of questions which have been shown in user testing to support these goals. Since the main interview questions and their order are mostly fixed, dialogue management concentrates on providing appropriate verbal feedback behaviors to keep the user engaged, maintain a natural and comfortable conversation flow, and elicit continuations and elaborations from the user.

The agent is implemented using a modular architecture (Hartholt et al., 2013). Dialogue processing, which is the focus of this demonstration, is supported by individual modules for speech recognition, language understanding and dialogue management (see Section 2). The agent’s language and speech are executed by selecting from pre-recorded audio clips. Additional agent modules include nonverbal behavior generation, which matches appropriately timed body movements to the agent’s speech; character animation in a virtual 3D environment; and rendering in a game en-

gine. The perception system analyzes audio and video in real time to identify features such as head position, gaze direction, smile intensity, and voice quality. DeVault et al. (2014) provides details on all the agent’s modules.

2 Overview of Dialogue Processing

2.1 ASR and NLU components

Unlike many task-oriented dialogue domains, interview dialogues between SimSensei Kiosk and participants are naturally open-ended. People tend to respond to interview stimuli such as “*what’s one of your most memorable experiences?*” with idiosyncratic stories and events from their lives. The variability in the vocabulary and content of participants’ answers to such questions is so large that it makes the ASR task very challenging. Furthermore, continuous ASR is employed to ensure that participants feel comfortable interacting with the system without being distracted by having to use a push-to-talk microphone. The use of continuous ASR necessitates the development of specific policies for turn-taking (see Section 2.2). In this demonstration, voice activity detection and speech recognition are performed using Pocket-Sphinx (Huggins-Daines et al., 2006).

Because of the open-ended participants’ responses, for NLU, we cannot simply construct a small semantic ontology and expect to cover the majority of meanings that will be expressed by users. Thus, this is an application in which the dialogue policy needs to be able to create a sense of engagement, continuation, and empathy despite relatively shallow and limited understanding of user speech. SimSensei Kiosk currently uses 4 statistically trained utterance classifiers to capture different aspects of user utterance meaning.

The first NLU classifier identifies generic dialogue act types, including statements, yes-no questions, wh-questions, yes and no answers, and several others. This classifier is trained using the Switchboard DAMSL corpus (Jurafsky et al., 1997) using a maximum entropy model.

The second NLU classifier assigns positive, negative, or neutral valence to utterances, in order to guide Ellie’s expression of empathy. We use SentiWordNet 3.0 (Baccianella et al., 2010), a lexical sentiment dictionary, to assign valence to individual words spoken by users (as recognized by the ASR); the valence assigned to an utterance is based primarily on the mean valence scores of

Opening Rapport Building Phase	
Ellie	<i>What are some things you really like about LA?</i> (top level question)
User	<i>I love the weather, I love the palm trees, I love the beaches, there’s a lot to do here.</i>
Diagnostic Phase	
Ellie	<i>Have you noticed any changes in your behavior or thoughts lately?</i> (top level question)
User	<i>Yes.</i>
Ellie	<i>Can you tell me about that?</i> (continuation prompt)
User	<i>I’m having a lot more nightmares now uh can’t sleep have haven’t really been eating uh trying to eat... I have to force down food um just feeling like an emotional wreck.</i>
Ellie	<i>I’m sorry to hear that.</i> (empathy response)
Ellie	<i>What are you like when you don’t sleep well?</i> (follow-up question)
User	<i>Irritable, emotional, it just adds to my overall stress um [long pause]</i>
Ellie	<i>What...</i> (Ellie speaks after the participant’s long pause)
User	<i>Can’t concentrate uh I uh...</i> (the participant starts speaking while Ellie is speaking)
Ellie	<i>I’m sorry please continue.</i> (Ellie realizes that she has interrupted the participant and apologizes)
Positive Closing Phase	
Ellie	<i>How would your best friend describe you?</i> (top level question)
User	<i>As caring, as fun because most of the time when I’m around my best friends I’m happy and I’m fun loving. I joke around with them a lot and uh I do better when I’m around my friends...</i>

Figure 2: Examples of Ellie’s interview phases.

the individual words in the utterance.

The third NLU classifier supports domain-specific small talk by recognizing a handful of specific anticipated responses to Ellie’s rapport-building questions. For example, when Ellie asks users where they are from, this classifier detects the names of commonly mentioned cities and regions using keyphrase spotting.

The fourth NLU classifier identifies domain-specific dialogue acts, and supports Ellie’s follow-up responses to specific questions, such as “*how close are you to your family?*”. This maximum entropy classifier is trained using face-to-face and Wizard-of-Oz data to detect specific responses such as assertions of closeness.

2.2 Dialogue Management

Ellie currently uses about 100 fixed utterances in total in the automated system. She employs 60 top level interview questions (e.g., “*do you travel a*

lot?”), plus some follow-up questions (e.g., “*what do you enjoy about traveling?*”) and a range of backchannels (e.g., “*uh huh*”), empathy responses (e.g., “*that’s great*”, “*I’m sorry*”), and continuation prompts (e.g., “*tell me more about that*”).

The dialogue policy is implemented using the FLoReS dialogue manager (Morbini et al., 2012). The policy groups interview questions into three phases (opening rapport building, diagnostic, positive closing – ensuring that the participant leaves with positive feelings). Questions are generally asked in a fixed order, with some branching based on responses to specific questions.

Rule-based subpolicies determine what Ellie’s follow-up responses will be for each of her top-level interview questions. The rules for follow-ups are defined in relation to the four NLU classifiers and the duration of user speech (measured in seconds). These rules trigger continuation prompts and empathy responses under specific conditions.

The turn-taking policy supports our design goal to encourage users to openly share information and to speak at length in response to Ellie’s open-ended questions. In this domain, users often pause before or during their responses to think about their answers to Ellie’s personal questions. The turn-taking policy is designed to provide ample time for users to consider their responses, and to let users take and keep the initiative as much as possible. Ellie’s turn-taking decisions are based on thresholds for user pause duration, i.e., how much time the system should wait after the user has stopped speaking before Ellie starts speaking. These thresholds are tuned to the face-to-face and Wizard-of-Oz data to minimize Ellie’s interruption rate, and are extended dynamically when Ellie detects that she has interrupted the participant. This is to take into account the fact that some people tend to use longer pauses than others.

Examples of the three interview phases and of Ellie’s subdialogue policies (top level and follow-up questions, continuation prompts, empathy responses, and turn-taking) are given in Figure 2.

3 Demonstration Summary

The demonstration will feature a live interaction between Ellie and a participant, showing Ellie’s real-time understanding and consequent policy actions. Live dialogues will highlight Ellie’s strategies for questioning, follow-up continuation prompts, displays of empathy, and turn-taking,

similar to the example in Figure 2. The demonstration will illustrate how these elements work together to enable Ellie to carry out extended interviews that also provide information relevant to the automatic assessment of indicators of distress.

Acknowledgments

The effort described here is supported by DARPA under contract W911NF-04-D-0005 and the U.S. Army. Any opinion, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

- S. Baccianella, A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of LREC*.
- D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of SIGDIAL*.
- D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of AAMAS*.
- A. Hartholt, D. Traum, S. Marsella, A. Shapiro, G. Stratou, A. Leuski, L.-P. Morency, and J. Gratch. 2013. All together now, introducing the virtual human toolkit. In *Proceedings of IVA*.
- D. Huggins-Daines, M. Kumar, A. Chan, A.W. Black, M. Ravishankar, and A.I. Rudnicky. 2006. Pocket-Sphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Proceedings of ICASSP*.
- D. Jurafsky, E. Shriberg, and D. Biasca. 1997. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13.
- F. Morbini, D. DeVault, K. Sagae, J. Gerten, A. Nazarian, and D. Traum. 2012. FLoReS: A forward looking reward seeking dialogue manager. In *Proceedings of IWSDS*.
- S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. 2013. Automatic behavior descriptors for psychological disorder analysis. In *Proceedings of IEEE Conference on Automatic Face and Gesture Recognition*.