‘How was your day?’ An affective companion ECA prototype

Marc Cavazza
School of Computing
Teesside University
Middlesbrough TS1 3BA
m.o.cavazza@tees.ac.uk

Raúl Santos de la Cámara
Telefónica I+D
C/ Emilio Vargas 6
28043 Madrid
e.rsai@tid.es

Markku Turunen
University of Tampere
Kanslerinrinne 1
FI-33014
mturunen@cs.uta.fi

José Relaño Gil
Telefónica I+D
C/ Emilio Vargas 6
28043 Madrid
joserg@tid.es

Jaakko Hakulinen
University of Tampere
Kanslerinrinne 1
FI-33014
jh@cs.uta.fi

Nigel Crook
Oxford University
Computing Laboratory
Oxford OX1 3QD
nigc@comlab.ox.ac.uk

Debora Field
Computer Science
Sheffield University
Sheffield S1 4DP
d.field@shef.ac.uk

Abstract

This paper presents a dialogue system in the form of an ECA that acts as a sociable and emotionally intelligent companion for the user. The system dialogue is not task-driven but is social conversation in which the user talks about his/her day at the office. During conversations the system monitors the emotional state of the user and uses that information to inform its dialogue turns. The system is able to respond to spoken interruptions by the user, for example, the user can interrupt to correct the system. The system is already fully implemented and aspects of actual output will be used to illustrate.

1 Introduction

Historically, Embodied Conversational Agents (ECAs) have been used in research and industry make information and complex tasks more accessible to customers and users. With the rise of new technologies in affective dialogue systems, we are beginning to see a future in which ECA dialogues are not all task-driven, but some will be focused on the social aspects of conversation. We envisage the development of ECAs that enhance the social quality of life of the user, and that build deep relationships with their users over years of use. Our affective companion ECA is our first contribution to this emerging technology.

Our emotionally intelligent dialogue system requires a high level of understanding of the meaning of the user turns, and so is quite unlike a ‘chatbot’. Affective aspects of the situation are also key in influencing the course of the conversation. The system attempts to balance both semantic and affective understanding in order to create an engaging and enjoyable companion for social conversation.

2 Scenario description

The system conversation is free-ranging and follows the user’s initiative, rather than being task-driven. We have for the time being restricted the range of conversation topics to things that happen in the life of someone who works in an office (the ‘How was your day?’ (HWYD) scenario). The user ‘offloads’ about his/her day at work by chatting to the ECA, and the ECA tries to empathise with and advise the user. The ECA asks questions to elicit information from the user, makes empathetic and sympathetic comments, and offers advice.

The system is able to process long user turns, as well as to generate long system turns (around 50 words). If the user perhaps dislikes what he/she hears, he/she can interrupt the system by barging in, and the ECA will respond in a human-like way. The conversation style is therefore considerably different from traditional task-based dialogues that consist of short dialogue turns and in which the system takes all the initiative.

The dialogue in Table 1 (actual output from a user session) illustrates some of the significant features of the system, including long system

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 277–280,
The University of Tokyo, September 24-25, 2010.

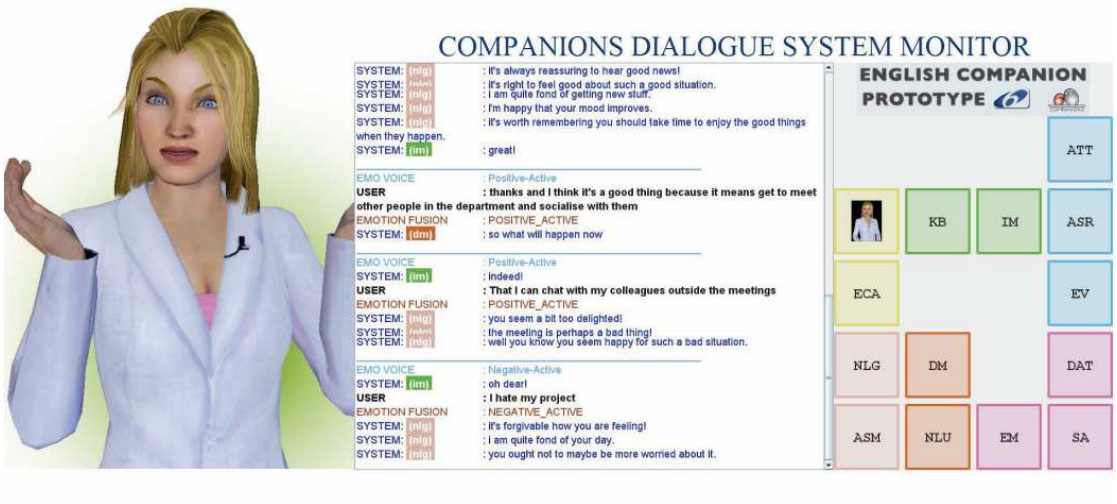

Figure 1: Screenshot of the prototype interface

turns, user interruptions and emotional analysis of user input.

User$_1$	Worst day ever. I arrived late because of the traffic.
System$_1$	You have my sympathy. What happened next?
User$_2$	I missed the meeting on project responsibilities; got more work dumped on me. Just what I don't need. I'm already likely to miss the deadline. [*ECA looks worried when user's negative tone is detected*]
System$_2$	Sorry to hear that. Who was at the meeting?
User$_3$	Only my team. They gave me a set of PowerPoint slides from the meeting. I spent the rest of the day reading them but my computer kept crashing because of updates.
System$_3$	Oh dear! It is understandable how you must be feeling. It's right to be unhappy about such a bad situation. Try not to worry about missing the deadline because of the technical problems. [*At this point user interrupts ECA by beginning turn User*$_4$] It's sad that your motivation could be affected by them. [*System stops speaking*]
User$_4$	No, I'm not worried about the technical problems, I'm worried about the amount of extra work on this project and the deadline.
System$_4$	What is the project about?

Table 1: Dialogue example

A video demonstration of the system in action is available at:
`http://www.youtube.com/watch?v=BmDMNguQUmM`

3 Architecture

Figure 1 shows a screen shot taken at run-time of actual system output. The ECA is represented on a screen as a woman (waist up) who displays natural, human-like movements and performs a wide range of complex facial expressions, bodily movements, and hand and arm gestures.

The screen also displays a transcript of the user and system turns. The user turns shown constitute the output of the Automatic Speech Recogniser (ASR). The system's analysis of the user's emotional state is also shown.

The right-most panel of the screen shows graphics which convey real-time information about how the dialogue is being processed. It presents a streamlined view of the software modules that comprise the system. Module activity is visually represented at run-time by flashing colours. This 'glass-box' approach enables detailed observation and analysis of system procedure at run-time.

The system comprises a number of distinct modules that are connected using Inamode, a hub-based message-passing framework using XML formatted messages over plain text sockets.

The system's ASR is the Nuance™ dictation engine. This is run in parallel with our own acoustic analysis pipeline which extracts low level (pitch, tone) speech features and also high-level features such as emotional characteristics. Analysis of the emotions is currently carried out

by EmoVoice (Vogt et al. (2008)). The ASR output strings are analysed for sentiment by the AFFECTiS system (Moilanen and Pulman (2007, 2009)) and classed as positive, neutral, or negative. This output is fused with the output from EmoVoice to generate a value that represents the user's current emotional state, which is expressed as a valence+arousal pairing (with five possible values).

The ASR output goes to our own Natural Language Understanding (NLU) module which performs syntactic and semantic analysis of user utterances and derives noun phrases and verb groups and associated arguments. Events relevant to the scenario (*e.g.*, promotions, redundancies, meetings, arguments, *etc*.) are recognised by the NLU and are used to populate an ontology (a model of the conversation content). The system is currently able to recognize and respond to more than 30 event types.

The events recognised in a user turn are labelled with the output of the Emotion Module for that turn; the result is a representation of both the semantic and affective information that the user might be trying to convey.

Our own rule-based Dialogue Manager (DM) takes the affect-annotated semantic output of the NLU, and from that and its model of the conversation content determines the next system turn. It will either ask a question about the events that occurred in the user's day, express an opinion on the events already described, or make empathetic comments. Whenever the system has gained sufficient understanding of a key event in the user's day, it generates a complex long turn that encapsulates comfort, opinion, warnings and advice to the user.

These long system turns are generated by our own plan-based Affective Strategy Module that makes an appraisal of the user's situation and generates an appropriate emotional strategy (Cavazza et al. (2010)). This strategy—expressed as an abstract, conceptual representation—is handed to our own Natural Language Generator (NLG) that maps it into a series of linguistic surface forms (usually 4 or 5 sentences). We use a style-controllable system using Tree-Furcating Grammars (an extension of the Tree-Adjoining Grammars formalism (Joshi et al. (1997)). This ensures the generation of a large set of different surface forms from the same semantic input.

The output of the NLG is passed to a module that adds this information to its system turn instructions for the ECA. The ECA has been developed around the Haptek™ toolkit and is controlled using an FML-like language (after Hernández et al. (2008)). This 2-D embodiment produces gestures, facial expressions, and body movements that convey the emotional state of the ECA. Its movements and expressions enable it to visually display interest and enjoyment in talking to the user, and to display empathy with the user. The speech synthesis module is our own emotion-focused extension of the Loquendo™ TTS system. It includes paralinguistic elements such as exclamations and laughter, and emotional prosody generation for negative and positive utterances.

4 Special procedural features

A significant processing design feature of the system is that there are two main processing loops from user input to system output; a 'long loop' which passes through all the components of the system; and a 'short loop' or 'feedback loop' which will now be discussed (the procedure already described in Section 3 is the long loop procedure).

4.1 Feedback loop

The feedback loop ('short loop') bypasses many linguistic components and generates immediate reactions to user activity. The main function of the short loop is maintain user engagement by preventing unnaturally long gaps of ECA inactivity. The feedback loop engages the acoustic analysis components, the TTS, and the ECA. It is responsible for the generation of real-time (< 500 ms) reactions in the ECA in response to the emotional state of the user. It attempts to align both verbal behaviour (backchannelling) and non-verbal behaviour (facial expressions, gestures, and general body language) to the emotions detected during most recent user turn. In order to achieve a reasonable level of realism, these system reactions to the perceived emotional state of the user need to be perceptibly instantaneous. Using this short feedback loop that bypasses many of the linguistic components ensures this.

The feedback loop is also occasionally used to make sympathetic comments immediately after the user stops speaking. These act as acknowledgements of the emotion expressed by the user. An example can be seen in the $System_2$ turn of the example dialogue in Table 1:

1.*"Sorry to hear that. Who was at the meeting?"*

Here, the first utterance was spoken by the system within a few tenths of a second after the end

of the previous user turn ($User_2$). The system tried to identify the user's emotion in the previous turn and then to behave linguistically and visually in an empathetic way. The actual sympathetic utterance was randomly chosen from a set of 'negative emotion utterances' (there are also 'positive' and 'neutral' sets).

The second half of the system turn in (1) was derived by the system's 'long loop'. It is a question which refers to a meeting that the user mentioned in the previous turn. This 'meeting' event has been heard by the ASR, understood by the NLU system, remembered by the DM, and is now referred to by an appropriate definite noun phrase in the output of the NLG.

The feedback and main loops run in parallel. However, the feedback loop generates its speech output almost immediately, giving time for the main dialogue loop to complete its more detailed analysis of the user's utterance.

4.2 Handling user interruptions

This system has a complex strategy for handling situations in which the user interrupts long system turns. The system's response to 'barge-in' user interruptions is overseen by the Interruption Manager (IM), which is alerted by the acoustic input modules whenever a genuine user interruption (as opposed to, say, a backchannel) is detected during a long system utterance. When alerted, the IM instructs the ECA to stop speaking when it reaches a natural stopping point in its current turn (usually the end of the current phrase). The user's interruption utterance is processed by the long loop. Its progress is tracked and controlled by the IM, for example, it makes sure that the linguistic modules know that the current utterance is an interruption, whic means it requires special treatment. The DM has a range of strategies for system recoveries from user interruptions, including different ways of continuing, replanning, and aborting. An example of a user interruption is shown in Table 1. The user interrupts the long system utterance in the $System_3$ turn. The system's response to the interruption is to stop the speech output from the ECA, abort the long system turn altogether, and instead to ask for more details about the project that the user has just mentioned during the interruption. (See (Crook et al. (2010)) for a more detailed description of the IM.)

Acknowledgements

This work was funded by Companions, a European Commission Sixth Framework Programme Information Society Technologies Integrated Project (IST-34434).

We would also like to thank the following people for their valuable contributions to the work presented here: Stephen Pulman, Ramon Granell, and Simon Dobnick (Oxford University), Johan Boye (KTH Stockholm), Cameron Smith and Daniel Charlton (Teesside University), Roger Moore, WeiWei Cheng and Lei Ye (University of Sheffield), Morena Danieli and Enrico Zovato (Loquendo).

References

Cavazza, M., Smith, C., Charlton, D., Crook, N., Boye, J., Pulman, S., Moilanen, K., Pizzi, D., Santos de la Camara, R., Turunen, M. 2010 *Persuasive Dialogue based on a Narrative Theory: an ECA Implementation*, Proc. of the 5th Int. Conf. on Persuasive Technology (Persuasive 2010), to appear 2010.

Crook, N., Smith, C., Cavazza, M., Pulman, S., Moore, R., and Boye, J. 2010 *Handling User Interruptions in an Embodied Conversational Agent* In proc. of AAMAS 2010.

Hernández, A., López, B., Pardo, D., Santos, R., Hernández, L., Relaño Gil, J. and Rodríguez, M.C. (2008) *Modular definition of multimodal ECA communication acts to improve dialogue robustness and depth of intention*. In: Heylen, D., Kopp, S., Marsella, S., Pelachaud, C., and Vilhjálmsson, H. (Eds.), AAMAS 2008 Workshop on Functional Markup Language.

Joshi, A.K. & Schabes, Y. (1997) Tree-adjoining Grammars. *Handbook of formal languages, vol. 3: Beyond Words*, Springer-Verlag New York, Inc., New York, NY, 1997.

Moilanen, K. and Pulman. S. (2009). Multi-entity Sentiment Scoring. *Proc. Recent Advances in Natural Language Processing* (RANLP 2009). September 14-16, Borovets, Bulgaria. pp. 258--263.

Moilanen, K. and Pulman. S. (2007). Sentiment Composition. *Proc. Recent Advances in Natural Language Processing* (RANLP 2007). September 27-29, Borovets, Bulgaria. pp. 378--382.

Vogt, T., André, E. and Bee, N. 2008. *EmoVoice – A framework for online recognition of emotions from voice. Proc. Workshop on Perception and Interactive Technologies for Speech-Based Systems*, Springer, Kloster Irsee, Germany, (June 2008).