04121 Abstracts Collection
Evaluating Embodied Conversational Agents
— Dagstuhl Seminar —

Zsofia Ruttkay[1], Elisabeth André[2], Kristina Höök[3], W. Lewis Johnson[4] and Catherine Pelachaud[5]

[1] CWI, Amsterdam, NL
zsofi@cs.utwente.nl
[2] Univ. Augsburg, DE
[3] IT-University Kista, SE
[4] USC Marina del Rey, US
johnson@isi.edu
[5] Univ. Paris, FR
c.pelachaud@iut.univ-paris8.fr

Abstract. From 14.03.04 to 19.03.04, the Dagstuhl Seminar 04121 "Evaluating Embodied Conversational Agents" was held in the International Conference and Research Center (IBFI), Schloss Dagstuhl. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Critical evaluation of some implemented EcAs, issues and framework for evaluation and design

04121 Working Group 1 – ECAs and human-human interaction

A number of approaches to modeling ECA behaviors are based on a direct simulation of human behaviors. Consequently, it comes as no surprise that the use of data-driven approaches which allow us to validate design choices empirically has become increasingly popular in the ECA field. The trend already started in 1999 with a Dagstuhl Seminar on Multimodality where a working group on multimodal corpora was established. Since then, a number of useful tools have been developed which facilitate the annotation and analysis of multimodal corpora and boosted progress in our field. Building upon this experience, the objective of the present working group was to investigate the potential benefits of corpora for the creation and evaluation of ECAs.

Given the fact that excellent material from previous workshop on multimodal corpora was already available, our working group was not forced to spend much

time and effort on a survey of the current state of the art, but could concentrate on an update of a document on multimodal corpora instead. The document was provided by Jean-Claude Martin, and we decided to make it publicly available.

In addition, we designed a questionnaire for the Dagstuhl attendees in order to get a clearer picture about current trends in multimodal corpora from a representative group of ECA researchers. In the following, we summarize some of the results:

• Employed information resources As it turned out, ECA researchers rely on a large variety of resources to inform the design of their ECAs including recordings of users in "natural" or staged situations, TV shows, Wizard of Oz studies, movies, games and motion capturing data.

• Use of corpora Only around 60% of the Dagstuhl attendees make use of a corpus most of them relying on an annotation tool. The evaluation of the questionnaires also revealed that a surprisingly high number of different annotation tools and home-made annotation schemes are currently being used.

• Major problems with corpora Major problems we identified when analyzing the questionnaires were the limited ways to re-use corpora which are in most cases collected for a specific purpose. Furthermore, the creation of a model or the extraction of ECA behaviors from a corpus is still an open research question. Also there is the danger, that human users expect a different behavior from an ECA than from a human conversational partner which might limit the potential benefits of a simulation-based approach.

• Criteria for the evaluation of ECAs All participants indicated that they performed a variety of experiments to analyze the user's objective and/or subjective response to the ECA. About 50% came up with a catalogue of evaluation criteria. Only 25% based their evaluation on a comparison of the ECA's behavior with that of human conversational partners.

Our working group agreed upon that the use of a corpus provides a promising approach to the modeling of ECA behaviors since it allows us to ground ECA behaviors in empirical data. In addition, we identified some interesting new options. For instance, we discussed to extract ECA behaviors from cartoons in order to capture implicit knowledge from professional designers. Furthermore, a corpus might help to compare the behavior of different ECAs by serving as a kind of reference system. In turn, an ECA might be employed to validate corpus annotations, e.g. by visualizing certain extracted behaviors. A comparison of the visualized and the original behaviors might then serve as a measurement for the accuracy and completeness of the corpus annotations.

As a common future project for ECA researchers interested in corpora we discussed the realization of culture-specific ECA's based on the recording and analysis of standardized staged or natural situations (e.g. asking for directions in different countries).

Joint work of: André, Elisabeth; Martin, Jean-Clude; Otero, Nuno; Rehm, Matthias; Ruttkay, Zsofia

04121 Working Group 2 – ECA's design parameters and aspects

How does one go about designing a human? With the rise in recent years of virtual humans this is no longer purely a philosophical question. Virtual humans are intelligent agents with a body, often a human-like graphical body, that interact verbally and non-verbally with human users on a variety of tasks and applications. Our working group approached this question from the perspective of interactivity. Specifically, how can one design effective interactive experiences involving a virtual human, and what constraints does this goal place on the form and function of an embodied conversational agent.

Our group grappled with several related questions: What ideals should designers aspire to, what sources of theory and data will best lead to this goal and what methodologies can inform and validate the design process? A longer article (.pdf) summarizes the output of this WG and suggests a specific framework, borrowed from interactive media design, as a vehicle for advancing the state of interactive experiences with virtual humans.

Joint work of: Gratch, Jonathan; Egges, Arjan; Eliens, Anton; Isbister, Katherine; Marsella, Stacy; Paiva, Ana; Rist, Thomas; ten Hagen, Paul

04121 Working Group 3 – Micro-level evaluation of ECAs(single modalities and aspects)

Embodied conversational agents (ECAs) implement all kinds of (para)linguistic and behavioral models. Typically, these are included for a particular purpose, for instance, to manage the interaction between a user human and the ECA (e.g., using gaze) or to suggest a particular emotion (e.g., by lowering the eyebrows). Micro-evaluation is a method to test whether the way these behaviors are implemented in the ECA is understood by users in the intended. Arguably, this evaluation is primary, since the added value of ECAs in applications might not be provable before being sure that the underlying models (and their implementation) are correct.

In this working group, we discussed micro-evaluation of:
(1) audio-visual speech,
(2) non-verbal behavior,
(3) natural language content,
(4) dialogue control and interaction and
(5) personality and emotion.

The micro-evaluation paradigm in these respective subfields appears to follow a general pattern. As a starting point, an ECA developer can look for models in the literature. As it turns out, many relevant models have been developed and published (in phonetics, conversational analysis, cognitive science, social psychology etc.). These models are often taken as a starting point and implemented in the ECA. Via judgment (perception) studies, it can be checked whether the

ECA implementation is faithful to the model. However, these models are often incomplete from an ECA perspective, for instance because they often lack information about the timing and execution of specific behavioral aspects. If that happens to be the case, additional effort has to be undertaken to fill in the missing pieces in a particular model. This can be done, for instance, using elicitation (production) studies with human speakers. The collected data can either be used to fine-tune the model or as part of a data-driven approach to ECAs.

While discussing the five topics listed above, it was found that the models get less detailed and more controversial when we move to higher-level issues. Thus, while audio-visual speech and non-verbal behavior are relatively well-understood, the models for personality and emotion are less easy to apply to ECAs. In these fields, more 'foundational' work is needed. Another general observation that came up at various times is that setting up good perception studies is no easy matter. It does not seem to be a good strategy to ask subjects directly whether they feel the ECA under evaluation has property X (e.g. is trustworthy), since the results of such experiments tend to be somewhat unreliable. Rather the perception test has to be set up in such a way that subjects have to make functional use of the ECA and thereby may show indirectly whether the agent has property X. This is what makes micro-evaluation difficult, but also what makes it so much fun to do.

Joint work of: Krahmer, Emiel; Beskow, Jonas; Cassell, Justine; Heijlen, Dirk; Marriott, Andrew; Masaro, Dominic; Noot, Han; Olivier, Patrick; Pele, Danielle

04121 Working Group 4 – System-level evaluation of ECAs(ECA in application, usage context)

ECAs are embedded in applications that interact with users, in some task and environmental context. In order to develop a clear picture of ECA performance, it is necessary to consider the user(s) and the application as a system, and evaluation the overall performance of this system. One must consider what environmental factors might have an impact on the performance of this system. This can be helpful in interpreting the role of micro-level ECA evaluationin determining system-level performance; one must make sure that micro-level evaluations are performed in a context that is comparable to the context of the user-application system.

System-level evaluation ideally involves three things: 1) evaluation of user-ECA performance, e.g, how fluent and efficient it is, 2) evaluation of the user experience, from a subjective standpoint, and 3) evaluation of the effectiveness of the ECA-enabled application in achieving its goals, e.g., learning outcomes or entertainment engagement. Each of these contributes on our understanding of the role of the ECA in the system.

Relevant environmental factors may include target user group, physical environment of use, integration into a larger work activity, stance of the user toward

the environment (observer or participant), etc. The role of the ECA in the application may be central or ancillary, and may change over time; the ECA may play the role of personal assistant, companion, antagonist, advisor, tutor, etc.

ECA development ideally involves a spiral approach, and evaluation should be incorporated into that spiral as well. This means for example that evaluations need to be simple enough and easy enough to perform to inform the ongoing design of the ECA, but at the same time provide information that is predictive of what the ultimate system-level performance will be.

Evaluation methods can involve a combination of extended observations, interviews, and questionnaires. Logs of interaction data, physiological data, and videotaps can all be helpful.

The system-level view can offer a different perspective on questions that are commonly asked regarding ECAs. For example, believability is a concern for ECAs, and believability is often assessed by having subjects observe ECAs and give their judgments. But passive observers may get a different impression of ECAs than users who are engaged in interacting with the ECA. It would be better to engage users in an interaction with the ECA, and then assess both the user's subjective impression of believability and system performance characteristics that correlate with believability.

Joint work of: Johnson, Lewis; Barker, Tim; Bernsen, Niels Ole; Biswas, Gautam; Cavazza, Marc; Noor, Christoph; Gerhard, Michael; Nijholt, Anton; Prendinger, Helmut

04121 Working Group 5 – Sharing work and results

One of the major goal in this group was to make an arsenal of models, tools and resources to facilitate ECA development and evaluation. We set up a site for ECA-tools, which hopefully will grow further. Please contribute, contact Igor Pandzic.

Another way of comparing and sharing ECAs is a kind of show or festival where ECAs in different application categories (like presenter, tutor) are compared, possibly on benchmark presentation tasks. It is not easy to come up with a framework where different aspects of ECAs can be fairly compared. Note that we are not aiming at a kind of 'Miss/Mr ECA' competition, but an ECA-challenge event where inventive designs, quality of communicational capabilities, novelty of applications can be reviewed. A handful of enthusiastic experts are busy with getting the ball rolling Ũ more help and suggestions are welcome, contact zsofi@cs.utwente.nl.

Initially we intended also to produce a uniform terminology and definition for design and evaluation aspects. The idea was partially addressed in the working groups 2-4, but there was not enough time for a joint and careful investigations. This is an item left for the future work.

Joint work of: Pelachaud, Catherine; Ishizuka, Mitsuru; Pandzic, Igor; Ruttkay, Zsofia

Ethical Dimensions of Synthetic Personalities

Tim Barker (Staffordshire University - Stafford, GB)

Do we want our characters displaying sexist, racist, violent behaviour ? Do we need to consider ethical issues to evaluate whether or not our characters are effective ?

Keywords: Synthetic personalities, ethics

Issues in Evaluating Conversation with (Virtual) M.C. Andersen

Niels Ole Bernsen (Univ. of Southern Denmark - Odense, DK)

The slides briefly summarise the status of our work on evaluating domain-oriented conversation with Andersen.

Keywords: Domain-oriented conversation, embodied conversational agents

Evaluating Teachable Agents: Educational Software that implements the Learning by Teaching Paradigm

Gautam Biswas (Vanderbilt University, USA)

The idea that teaching others is a powerful way to learn is both intuitively compelling, and one that has garnered support in the research literature. The literature on tutoring suggests a similar conclusion in that tutors have been shown to benefit as much from tutoring as their tutees Biswas and colleagues (Biswas, Schwartz, Bransford, & TAG-V, 2001) report that students preparing to teach made statements about how the responsibility to teach forced them to gain deeper understanding of the materials. A key benefit of the learning by teaching process focuses on the need to structure knowledge in a compact and communicable format. This requires a level of abstraction that may help the teacher develop important explanatory structures for the domain. The need to structure ideas not only occurs in preparation for teaching, but can also occur when teaching to accommodate the learners existing knowledge structures. Good learners bring structure to a domain by asking the right questions to develop a systematic flow for their reasoning. Good teachers build on the learners' knowledge to organize information, and in the process, they find new knowledge organizations, and better ways for interpreting and using these organizations in problem solving tasks.

Reflection on these studies and others lead us to conjecture that the creation of embodied conversational agent environments, where students can assume the role of "teacher" and teach the agent, may provide an effective and motivating

environment for learning. We have designed a multi-agent environment called "Betty's Brain" in which students learn by explicitly teaching computer agents to solve problems and answer questions using causal map structures in scientific domains. Once taught, the teachable agent reasons with its knowledge and answers questions. Students observe the effects of their teaching by analyzing these responses, and through feedback from a mentor agent. Betty uses a multi-modal interface, with text, speech, and animation to interact with the student. Overall, the system combines learning by teaching with self-regulation strategies to promote deep learning and understanding. Scaffolds, to help novice learners, are provided in the form of hypertext resources and a Mentor agent. Two studies demonstrate the effectiveness of this system. The first study focused on components that define student-teacher interactions in the learning by teaching task. The second study conducted after self-regulation strategies were added to the computer environment compared tutoring, learning by teaching, and learning by teaching with self-regulation strategies.

In our work, we have found that evaluation of the agent has to be performed as much in terms of how much students learn while using the system as much as what they think of the visual interfaces and modes of interaction. We have used two forms of images to represent Betty: (i) a cartoon face, and (ii) a human face that expresses appropriate emotions (sad, happy, confused, etc.) when interacting with the student. However, this does not seem to have had significant impact. The primary issues that have influenced Betty are: (i) the clarity of the animation that Betty uses to explain answers to questions that she is asked, (ii) her responsiveness and interactions during the teach phase. Students want her to be more involved, and participate in the learning process by asking questions and making comments as she is being taught. We have been able to achieve this with self-regulation strategies, (iii) the nature of the scaffolding provided to the novice learners and teachers. This has been done through the use of better online resources, better organization of quiz questions, and more generic help from the Mentor agent as opposed to answering a quiz.

In summary, I will be very interested in studying and discussing how the application domain and the target audience for ECAs influence evaluation procedures? Is there a comprehensive evaluation framework that drives this process, or is it very application specific?

Keywords: Learning by teaching, multi-agent architecture, multi-modal teachable agent, evaluating learning gains

Joint work of: Biswas, Gautam; Schwartz, Daniel; and the Teachable Agents Group at Vanderbilt (TAG-V)

The Ones that Got Away: Evaluating ECAs

Justine Cassell (NW University - Evanston, USA)

For the last 8 years, my students and I have continued development on the same ECA platform, repeatedly adding interactional capabilities. In each case, innovation followed the same development cycle: (1) choosing a conversational/ discourse/ nonverbal phenomenon in human-human interaction, (2) collecting data on that phenomenon in human-human interaction, (3) analyzing those data to get a distribution analysis of the phenomenon, (4) building a formal model from the results of the analysis, (5) implementing an ECA based on the model, (6) finding a domain of application, and finally (7) evaluating human performance with the ECA by comparing human interaction with one version of the ECA that instantiates the phenomenon to interaction with a version that does not instantiate that phenomenon. This extensive experience with ECA design, implementation and evaluation – not all of it leading to positive outcomes – allows me to talk about the lessons learned both from our failures and our successes – from the phenomena that improved human performance or preference for the ECA, to those that had no effect, to those that made the user *disprefer* that version!

Realistic Body Motion Synthesis for Virtual Humans

Arjan Egges (Université de Genève, CH)

In this talk, we propose a novel animation approach based on Principal Component Analysis allowing to generate two layers of subtle motions: small posture variations and personalised change of balance. Such a motion generator is needed in many cases when one attempts to create an animation sequence out of a set of existing clips. In nature there exists no motionless character, while in computer animation we often encounter cases where no planned actions, such as waiting for another actor finishing his/her part, is implemented as a stop/frozen animation.

We identify many situations where a flexible idle motion generator can help: from synchronisation of speech/body animation duration, to dynamic creation of stand still variations in between two active plays. Our approach overcomes the limitations of using a small set of existing clips as a basis for synthesizing idle motions, such as unnatural repetition of movements and difficulties to insert idle motions into an animation without breaking its continuity. A realistic animation is obtained by blending the small posture variations and the personalised balance shifting animations.

Keywords: Body Animation, Gesture Synthesis, Virtual Characters

Joint work of: Egges, Arjan; Magnenat-Thalmann, Nadia

ECA Perspectives - Requirements, Applications, Technology

Anton Eliens (Vrije Universiteit Amsterdam, NL)

In the last years we have developed a platform for the realization of embodied (conversational) agents, in a distributed logic programming framework. In this paper we will present an overview of our work, by discussing the requirements that acted as our guidelines for design decisions during development, some of the applications that have served as target demonstrators for developing and testing new functionality, and the (distributed logic programming) technology which we used for the realization of the platform and the implementation of our STEP scripting language.

Although the focus of our paper will primarily be our own DLP+X3D platform, we believe that our discussion along the perspectives of requirements, applications and technology might be more generally worthwhile in establishing the relative merits of the operational use of ECA-technology. At the end of this paper, we will moreover provide some hints of how to approach the experimental validation of the (possible) benefits of embodied conversational agents in user applications.

Keywords: Embodied agents, virtual environments, rich media

Joint work of: Eliens, Anton; Huang, Zhisheng.; Hoorn, Johan F.; Visser, Cees T.

Full Paper: http://drops.dagstuhl.de/opus/volltexte/2006/461

Evaluating Embodied Conversational Agents in Collaborative Virtual Environments

Michael Gerhard (Fraunhofer ISST, D)

There are currently no evaluation methods specific to ECAs in CVEs and traditional evaluation methods are limited in their applicability and consequently unlikely to address the full range of aspects now inherent in such systems. We argue that a combination of controlled experimentation, quasi-experiments, review-based evaluation and heuristic expert reviews is needed. To operationalise these traditional evaluation methods the concept of presence was deployed, and it was argued that presence as a cognitive variable can be measured and that such a measure can be a key indicator of the usability of ECAs in CVEs. Presence measures can be administered within controlled experiments and quasi-experiments to test certain aspects of the system. Such experiments might turn out particularly useful as a means of selecting between two or more design options and it is argued that issues concerning ECAs in CVEs can be meaningfully evaluated by comparing subjects' experience of presence. Further, although implementation

issues were not the primary concern of this study, the strength and shortcomings of the prototype agent were evaluated as secondary variables within that experiment. A set of criteria developed for the evaluation of the strengths and shortcomings of the current prototype agent are partly based on Nielsen's general usability guidelines and partly on a set of heuristics proposed for non-embodied conversational systems.

Keywords: Embodied Conversational Agents,Collaborative Virtual Environments, Presence

Extended Abstract: http://drops.dagstuhl.de/opus/volltexte/2006/460

Virtual Humans with Emotion

Jonathan Gratch (USC - Marina del Rey, USA)

My talk gives an overview of our work on virtual humans in the mission rehearsal exercise project, and our work on computational models of emotion (or more specifically, appraisal theory) and how they inform the verbal and non-verbal behavior of these interactive characters

Keywords: Virtual humans, emotion, appraisal theory

MPML-Mobile for describing ECA contents on Mobile Phones

Mitsuru Ishizuka (University of Tokyo, J)

We have been working on multimodal interfaces & contents with lifelike characters, focusing on our XML-based description language called MPML (Multimodal Presentation Markup Language).

One emphasis on MPML is easy description for anyone (ordinary people, not animation experts). Like HTML for current Web contents, we expect MPML to serve as an information-conveying vehicle in the Internet.

The main target of MPML is Web contents on PCs; however, we have extended it to other information environments while keeping its basic style.

In my talk, I introduced the following items with some demonstrations. 1) Several versions of MPML with their functions and tools. 2) An emotion-rich character agent SmArt (a talking head). 3) MPML-Mobile version for mobile-phone contents with character agents. (This is going to be in a commercial (pay) service from the 2nd largest mobile-phone company in Japan shortly.) 4) MPML-HR for controlling Humanoid-Robot presentations (preliminary stage).

Evaluations of Cooperation between Modalities in ECAs

Jean-Cluade Martin ((LIMSI - CNRS, F)

Multimodal Interfaces involve several levels of abstraction. In the case of multimodal input interfaces, the system has to infer the userŠs goal on the basis of monomodal signals such as her speech and her gestures. In the case of ECA multimodal output interfaces, the system has to find out the adequate monomodal signals such as the speech and gesture of a virtual character depending on her goal or emotional state. My claim is that there is an intermediate level which can be helpful to model these relations which is based on a typology of cooperation or combination between modalities. I start by describing an experiment on the input modalities. In a conversational game settings we recorded the spoken and multimodal behavior of adults and children while interacting with 2D cartoon agents simulated by a hidden experimenter. Multimodal scenarios were significantly shorter And yield higher and more homogeneous ratings of easiness between adults and children. For multimodal scenarios, we studied the use of speech, pen, and their overlap (simultaneous use). Pen proved to be the interaction mode the most used. This main effect is due to childrenŠs behavior who gesture more than adults.

The second experiment is on the output modalities. We presented 3 multimodal strategies (redundant, complementary, speech only), 3 looks of agents for 3 pedagogical presentations. These combinations were counterbalanced across subjects. The analysis of the postexperimental questionnaire revealed the following results. Redundant and complementarity were rated higher than the speech-only. This was due to an interaction with gender : there is an effect of strategies for male subjects but not for female subjects.

Keywords: Multimodal strategies, 2D agents, evaluation, bidirectional interaction

Joint work of: Martin, Jean-Claude; Buisine Stéphanie

Evaluating Embodied Conversational Agents (ECAs) for Quality and for Effectiveness in Language Learning

Dominic W. Massaro (Univ. California - Santa Cruz, USA)

There are many reasons such as large individual differences that make it difficult to evaluate the quality and effectiveness of embodied conversational agents (ECAs). Several experimental techniques are available to assess their effectiveness for each individual, however. We illustrate three different techniques in the context of the quality of ECAs and their effectiveness for language learning. In the first technique, the quality of the ECA is compared to no ECA and to a real person. Examples from speech intelligibility and emotional quality can be described. The second involves the assessment of the influence of several sources of

information on a behavior outcome. The third technique involves an assessment of the effectiveness of an ECA for language learning under different experiment conditions. These techniques are illustrated in the context of several completed studies.

Short Talk about ECA Research in Twente

Anton Nijholt (University of Twente, NL)

Short talk about ECA Research in Twente

Keywords: Embodied conversational agents, educational environments, multi-party interaction

Evaluation of Animated Presentation Agents

Thomas Rist (DFKI Saarbrücken, D)

This talk is an introduction to the presentation by Elisabeth Andre. We introduce a number of selected sample systems that have been built in the past years as well as systems which are currently under development at DFKI.

Keywords: Animated Presenter, Embodied Conversational Agents