BCogS 2011 - Program

Workshop on Benchmarking and Evaluation of Interactive Cognitive Systems
Supported by Excellence Cluster Cognitive Interaction Technology (CITEC), Bielefeld University

Workshop program

Mo, Feb 7,2011

10:00 Reception
10:45 Opening of the workshop

11:00-12:30 First Session: History and Perspectives

Chair: Stefan Kopp

Sven Wachsmuth, "Evaluation of Interactive Cognitive Systems: Individual Experience or Systematic Benchmark?"
Minoru Asada, "How interactive and how cognitive can humanoids be with humans?"
Fabio Bonsignorio, "Refutable Robotics Research? Good Experimental Methodology and Benchmarking in Robotics Research"

12:30-14:00 Lunch

14:00-15:45 Second Session: Competitions

Chair: Agnes Swadzba

Tijn van der Zant, "Benchmarking Robot Cognition using Natural Language"
Alexander Koller, "Lessons from the Challenge on Generating Instructions in Virtual Environments (GIVE)"
Jim Little, "Evaluation via Competitions in Vision and Robotics"
Christina Unger, "Benchmarking question-answering systems"

15:45-16:15 Coffee break

16:15-17:35 Third Session: Tasks in the real world

Chair: Sven Wachsmuth

Martin Buss, "The Autonomous City Explorer ACE Robot"
Tamin Asfour, "The Robo-KITchen Humanoid Benchmark"
Jochen Steil, "Benchmarking of learning motion primitives"

17:40-19:00 Panel Discussion

Moderator: Stefan Kopp

Minoru Asada (U Osaka), Martin Buss (TUM), David Traum (USC), Vincent C. Müller (euCogII), Helge Ritter (UniBi)

20:00 (Invited) Dinner in "Glück und Seligkeit"

Di, Feb 8,2011

9:00-10:30 Fourth Session: Methodologies I

Chair: Manja Lohse

Martin Riedmiller, "Lessons learned from Competitions"
Friederike Eyssel, "(Measuring) Anthropomorphism: A social psychological perspective"
Jan-Peter de Ruiter, "The Turing Test Light - a behavioral way to evaluated interactive systems"

10:30-11:00 Coffee break

11:00-12:30 Fifth Session: Communicative behavior

Chair: Karola Pitsch

David House, "Evaluation of speech, prosody and multimodality in interactive systems: from listener tests to user behavior"
Stefan Kopp, "Data-based modeling and evaluation of conversational agents"
David Traum, "Evaluation of Dialogue Systems and Virtual Humans"

12:30-13:45 Lunch

13:45-15:15 Sixth Session: Systems and Design

Chair: Ulrich Rückert

Gorgio Metta, "Common platforms and benchmarks for cognitive system"
Rudolphe Gelin, "A humanoid robot for cognitive applications"
Sebastian Wrede, "Runtime Analysis of Robotics Software Architectures and Benchmarking"

15:15-15:45 Coffee break

15:45-17:15 Seventh Session: Methodologies II

Chair: Friederike Eyssel

Sebastian Möller, "Evaluation of Cognitive Interactive Systems: Problem Formulation and First Insights"
Manja Lohse, "Interaction-inspired system design based on benchmarks"
Astrid Weiss, "The USUS Evaluation Framework: How to assess usability, social acceptance, user experience, and societal impact in HRI" (cancelled)

17:15-17:30 Closing Remarks

Evaluation of speech, prosody and multimodality in interactive systems: from listener tests to user behavior.

David House (KTH, Stockholm)

In the traditional evaluation of the quality of speech synthesis used in interactive systems, speech has generally been seen as a free-standing component. Using listener testing paradigms, intelligibility and naturalness have been the main criteria of such evaluations which have evolved into large-scale challenges at major speech conferences during the past few years. The evaluation of audio-visual synthesis using talking heads has used similar testing paradigms. While this type of assessment has undoubtedly led to improvements in synthesis and important advances and exchanges in techniques, a shift from a listener judgment perspective to a user behavior perspective can have important implications for our understanding of how speech can be better exploited in interactive dialogue systems. One example of this is prosody. While evaluations of intelligibility and naturalness of speech synthesis necessarily involve prosody, testing the prosodic component as a separate signal of specific aspects of system cognition can be useful. By assessing user behavior with such measures as task completion, reaction time or interaction analysis, we are able to obtain a different type of evaluation of how speech and prosody work in an interactive system. In this presentation, I will present and discuss some of the large-scale challenges in audio and audio-visual speech synthesis presented at major speech conferences during the past few years. I will further present some of the past and current work at KTH involving evaluating various aspects of prosody and multimodality in interactive dialogue systems.

The USUS Evaluation Framework: How to assess usability, social acceptance, user experience, and societal impact in HRI

Astrid Weiss (University of Salzburg)

This talk presents a theoretical and methodological evaluation framework for the assessment of human-robot interaction with humanoid robots in terms of Usability, Social Acceptance, User Experience, and Societal Impact - the USUS evaluation framework. In the beginning the theoretical and methodological evaluation framework will be presented which is based on a multi level/ indicator approach and thus defienes all evaluation factors and indicators and the methods with which those can be addressed during the evaluation process. Afterwards an overview on application scenarios will be given, by the means of eleven case studies, conducted within the USUS framework. As a piece of exploratory research, the conducted case studies cannot guarantee a holistic validation of the theoretical and methodological evaluation framework, but the guideline-based application of the evaluation framework can positively influence the design process in Human-Robot Interaction projects.Thus, subsequently the feasibility and limitations of the framework will be discussed.

Lessons learned from Competitions

Martin Riedmiller (University of Freiburg)

In 1998, our team 'Brainstormers' entered the RoboCup soccer simulation world championship for the first time. It was our goal from the beginning to show, that machine learning methods can be successfully applied (and improved) in competitive settings, letting behind the usual 'proof-of-

concept' level. In our active years until 2009 we successfully participated in 4 different leagues, always trying to combine the scientific and the competitve aspect: 2D soccer simulation, 3D soccer simulation, MidSize league, and the Festo Logistics Competition. In 2005, I co-organized the first RL Benchmarking Competition at NIPS 2005. In this talk I will summarize some of our experiences participating in the competitions and will give a (personal) resume of these 11 years.

Evaluation of Cognitive Interactive Systems: Problem Formulation and First Insights

Sebastian Möller^1, Ina Wechsung^1, Christine Kühnel^1, Benjamin Weiss^1, Tim Polzehl^1, Felix Putze^2, Tanja Schultz^2

(1) Quality and Usability Lab, Deutsche Telekom Labs, TU Berlin, Germany

(2) Cognitive Systems Lab, Karlsruhe Institute of Technology, Germany

Whereas the assessment and evaluation of spoken dialog systems has reached a level of maturity which resulted in first standards, the evaluation of so-called “cognitive” interactive systems is still an open issue. Part of the problem is due to the fact that such systems communicate with the user in several modalities, making use of different types of sensor and actuator information. Another part is due to such systems sensing and adapting to the user, task, usage environment and location. As a consequence, the behaviour of the system differs in each interaction, making a controlled evaluation particularly difficult.

Fox example, we saw a considerable effort in user state (e.g. emotion) recognition in the last years. As emotions are of short-term nature, they obviously differ from situation to situation, which makes user and environmental influences dependent factors. In addition, spontaneous emotional reactions should be interpreted with the background of what the speaker’s personality is. Thus, the evaluation of interactive systems which consider the user state needs to be realized in a personality-dependent and situation-dependent way.

In this talk, we would like to draw a picture of the evaluation problem for cognitive interactive systems. We will start by identifying aspects of the Quality of Service (QoS) and the Quality of Experience (QoE) of such services. Whereas the first ones can be quantified in terms of performance metrics and interaction parameters, a quantification of the latter depends on the availability of adequate subjective metrics, such as questionnaires or physiological parameters. We will present some recently-developed metrics which have partially also been proposed as an ITU-T Supplement. On the other hand, we will identify the open problems with respect to sparse data in case of systems which adapt to their users and usage context. We will conclude with a set of research questions which have to be answered in order to make significant advances in resolving the evaluation problem for cognitive interactive systems.

Lessons from the Challenge on Generating Instructions in Virtual Environments (GIVE)

Alexander Koller (Saarland University)

joint work with Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander

Over the past few years, research challenges have become increasingly influential in the development of the field of natural language generation (NLG). One challenge that is particularly relevant for this workshop is the GIVE Challenge, in which an NLG system must generate real-time instructions that support a user in performing a task in a virtual 3D environment. By connecting NLG systems to users over the Internet, GIVE cheaply acquired task-based evaluation data from large numbers of experimental subjects; in the second challenge, which completed in 2010, from over 1800 users in total.

In my talk, I will first present the GIVE Challenge and highlight some details of its approach to NLG system evaluation. I will then discuss advantages and challenges that the GIVE evaluation approach faces. In particular, I will present some thoughts on what, exactly, GIVE can measure in terms of NLG quality and interactivity, and what this means for the definition of a "good NLG system".

Evaluation via Competitions in Vision and Robotics

Jim Little and David Meger (University of British Columbia)

A fundamental difference between a cognitive robot and an industrial robot, for example, is the range of tasks that can be accomplished, the flexibility of the robot in the face of novel variations in the task, and the generality of its solutions. Competitions and challenges should reveal cognitive advances and innovations, as well as relevant design decisions.

In practice, because robots must be tested in realistic situations, irrelevant aspects or contingencies determine the success of a robot on a challenge. Moreover, some situations are decided by engineering solutions tailored to the particulars of the challenge or scenario.

We will discuss existing challenges such as the Semantic Robot Vision Challenge and the recent Solutions in Perception Challenge (SPC) with the goal of identifying how to arrange the competitions so that the effects of standard platforms and software frameworks are minimized. One possible solution is to develop challenges as a series of successively more complex problem as in the SPC, or as collections of complementary tasks, as in Robotics at Home. We will examine how multi-modal corpora and testing scenarios, perhaps even in simulation, could play a role in future challenges, and speculate how even in simulation/replay robotic issues such as viewpoint selection can be addressed.

Benchmarking Robot Cognition using Natural Language

Tijn van der Zant (University of Groningen)

To benchmarking a robot's performance, it is essential to perform tests on the complete system level. Only benchmarking parts, such as algorithms or specific hardware components, do not always generalize when applied in a real robot that operates in the real world. Also, static benchmarking leads to overly specialized solutions. What is required is a dynamical/statistical approach towards

benchmarking. This has been applied in the RoboCup@Home competition since 2006 and the approach will be discussed briefly in the first part of the presentation.

Recently RoboCup@Home started with the first international benchmark on the cognitive aspects of robots, which results in even more questions regarding the correct approach and how to create international standards. In the second part of the presentation the 'General Purpose Service Robot' test will be explained, which is the cognitive test of RoboCup@Home. This test is the result of a 1.5 year discussion within the @Home community. In this test we actually talk with the robot for a while, upon which the robot demonstrates that it understands the human utterance by asking questions (if needed) and subsequently executing a task (or not executing it). The robot can get a complex command, an underspecified command or a confusing one. The test could serve as a template for more cognitive tests in other domains. During the discussion I would like to get feedback for improving the test for the coming years and get ideas to create more categories for this cognitive benchmark.

Common platforms and benchmarks for cognitive system

Gorgio Metta (Italian Institute of Technology)

I will make the case that cognitive systems and robotics have been recently benefiting from the appearance of various “common” platforms, being these hardware systems (now available more easily and at a lower price) and the software middleware to support such complex hardware [1]. Both have been helped by the adoption of Open Source licenses which enabled a frank and open confrontation of methods and approaches (much more than the traditional peer-review publications) as well as accumulation of knowledge (through sharing of the source code). There have been, for example, proponents of the inclusion of the source code in various forms in publications in order to enable others to check the results [2][3]. This is a clear step forward. It does not necessarily alleviate the problem of finding benchmarks for cognitive systems but at least it enables the possibility of constructing meaningful ones.

[1] ICRA 2009, Kobe Japan. SW-F8 Workshop on Open Source Software in Robotics.

[2] S. Fomel, J. F. Claerbout, Guest editors' introduction: Reproducible research, Comput. Sci. Eng. 11 (2009) 5-7.

[3] N. Barnes, Publish your computer code: it is good enough. Nature 467, 753 (2010) | doi:10.1038/467753a

The Autonomous City Explorer ACE Robot

Martin Buss (TU München)

The Autonomous City Explorer ACE robot is the first of its kind to be able to find its way by interacting with pedestrians. The challenge in itself and the successful system architecture are discussed. Pedestrian interaction and a goal-oriented cognitive system architecture are the key to implement autonomous and cognitive behavior of the robot in a complex outdoor environment.

A humanoid robot for cognitive applications

Rudolphe Gelin (Aldebaran Robotics)

Nao is a small humanoid robot developed by the French SME Aldebaran Robotics. It has been designed to be used by non robotic-expert. Thanks to very simple and efficient environment of development, it can be programmed for a wide spectrum of applications from navigation to cognitive interaction with human beings. More than 800 Nao are currently used by researchers all over the world. In the talk we will present Nao, its possible applications and the future developments.

How interactive and how cognitive can humanoids be with humans?

Minoru Asada (Osaka University)

In addition to the functional evaluation for interactive cognitive systems, structural evaluation is needed to understand such process which would help to makes such system more functional. Cognitive developmental approach to understand the process enables to suggest to design such systems. A case of the early vocal communication is shown, and argued.

Evaluation of Dialogue Systems and Virtual Humans

David Traum (University of Southern California)

Despite much work in the area, evaluation of the conversational capabilities of dialogue systems and virtual humans remains a challenging problem. Some of the main problems are:

- lack of objective performance metrics for all but the simplest domains.
- inappropriateness of "gold-standard" corpus techniques, given plurality of very different "good" contributions and subtle differences between good and bad contribuitions, as well as the contingent, context-dependent nature of dialogue such that a different contribution early on can render the rest of a collected dialogue irrelevant.
- difficulty generalizing across different kinds of dialogue genres.
- cross-system "challenges" tend to be focussed on a single domain or on "fooling" judges into believing the system is human.

In this talk I will review some of the state of the art techniques for evaluating dialogue systems in spite of these challenges, as well as some of our recent work to develop new evaluation methods, including qualitative human judgements, and evaluation understudies for automated learning of dialogue polcies.

Data-based modeling and evaluation of conversational agents

Stefan Kopp (Bielefeld University)

The attainment of conversational abilities, i.e. being able to engage in natural, situated face-to-face dialogue, is becoming increasingly important for interactive systems like collaborative agents or social robots. For the most part, technical approaches to modeling and processing conversational behavior like speech, gestures, facial expressions, or head movements have been proposed and implemented in this field based on theoretical conceptions and empirical data. However, rigorous and commonly acknowledged methods for evaluating and benchmarking such agents, their interactive capabilities and the underlying models are lacking. With the advent of more complete and deployable systems an increasing number user studies have been carried out, in which users interact with the system and rate it afterwards by filling in questionnaires. This method has proven useful to study social effects of the agents and robots, but it is of limited help in verifying or falsifying specific assumptions made in the modeling of their conversational abilities, even in controlled comparative experiments. I will discuss the ways in which the modeling of conversational behavior can and need to be evaluated and benchmarked, both during a data-based modeling process itself as well as using the final system. Concrete examples and lessons learned from the modeling of speech and gesture behavior are presented.

Interaction-inspired system design based on benchmarks

Manja Lohse (Bielefeld University)

Situations in which novice users come into contact with service robots are increasingly common. Millions of robots have already been sold, most of them being vacuum cleaners, lawn-mowing robots, and entertainment robots. Since these are restricted to simple tasks, they usually work quite well. For the future, much more complex scenarios such as robots for eldercare or helpers in the household are envisioned. These can only be successful if the developers know how the users interact with them and what expectations they have. Thus, system design needs to be inspired by interaction, or in other words by user-centered approaches which take human cognition, expectations, and behaviors into consideration. However, despite an interaction-inspired design process, in the past we repeatedly wondered how objective changes of the systems and their components were related to the users' perception. It sometimes happened that the objective performance of the robot improved but the users' evaluations of the systems in questionnaires and interviews remained the same or even got worse. In the talk, I want to introduce some explanations for this phenomenon and show how benchmarks might help to avoid it.

Refutable Robotics Research? Good Experimental Methodology and Benchmarking in Robotics Research

Fabio Bonsignorio (University Carlos III of Madrid)

As the complexity of current robotic and embodied intelligent systems grows, it is more and more necessary to define proper experimental approaches and benchmarking procedures and some properly epistemological questions naturally arise. On the one hand, reliable benchmarks are called for in order to allow the comparison of the many research results in robotics research, so that their

industrial application is eventually possible. On the other hand, if robotics aims to be regarded as serious science, replication of experiments deserves consciencious attention; it is necessary to be able to verify if and by which measure new procedures and algorithms proposed in research papers constitute a real advancement and can be used in new applications.

This poses several issues and suggest some innovation on the way research is usually performed and reported in our field.

(Measuring) Anthropomorphism: A social psychological perspective

Friederike Eyssel (Bielefeld University)

The talk proposes a social psychological perspective on anthropomorphism, resting on the 3-Factor-Model of Anthropomorphism by Epley, Waytz & Cacioppo (2007). In the presentation, the issue of how to measure anthropomorphism, i. e., the attribution of essentially human traits, characteristics or intentions, will be focal. What are - in essence - typically human features?

We believe that social psychological theorizing in the domain of intergroup discrimination research can provide valuable answers to this question.

Thus, we propose that measures derived from social psychological intergroup reseach on the derogation of human social groups can, in a theoretical inversion, be useful for research on nonhuman agents. We will show this in a series of experiments on key aspects that contribute to greater acceptance and anthropomorphism of both robots and virtual agents. Theoretical and practical implications of these findings for psychology and robotics will be discussed.

The Robo-KITchen Humanoid Benchmark

Tamim Asfour and Rüdiger Dillmann (Karlsruhe Institute of Technology (KIT)

Benchmarking cognitive systems research is a challenging and difficult task because of the high interdisciplinary nature of the field. In the past, different robotics competitions and benchmarks have been suggested, e.g. RoboCup Soccer, RoboCup Rescue, RoboCup@HOME, DARPA grand challenge, etc. Recently, the attention of the community at the intersection of robotics, cognitive systems and artificial intelligence has been increased to the definition of clear and reproducible evaluation performance metrics and benchmarking procedures in terms of methodology and test data sets, which are both representative for the involved disciplines and have the chance to be accepted by a wide community.

In this talk, we present our kitchen scenario as an example for benchmarking humanoid robots acting and interacting in human-centered environments and discuss cognitive challenges for humanoid robotics towards the development of metrics and benchmarks, which highlight and focus on the evaluation of behaviours and results at system level rather than focusing on the performance of single algorithms. We argue that benchmarks should be measurable and scalable at system level in open ended scenarios under changing conditions, and that humanoid platforms provide the most suitable test-bed to study benchmarking in cognitive systems.

Benchmarking Question Answering Systems

Christina Unger (Bielefeld University)

An important challenge in the context of the Semantic Web is to provide casual end users with easy access to the growing body of semantic data. Studies showed that users prefer search paradigms based on natural language question answering, that allow users to express arbitrarily complex information needs in an intuitive fashion. The essential problem lies in translating these information needs into a form such that they can be evaluated using standard Semantic Web query processing and inferencing techniques. I will present a shared task that we developed in order to evaluate and compare question answering systems, focusing on some of the difficulties in setting up a balanced gold standard.

Evaluation of Interactive Cognitive Systems: Individual Experience or Systematic Benchmark?

Sven Wachsmuth (Bielefeld University)

Evaluation and benchmarking is typically thought about in terms of assessing some kind of "objective" numbers that characterize the performance of system capabilities, that make a comparison of different systems possible, and that enable a judgement that system A is better than system B. For interactive cognitive systems like conversational agents or robots, this seems to be problematic as the research community is struggeling with these topics for quite a long time. The reasons for this are multifaceted and range from technical and conceptual issues to organisational issues.

In my talk, I will reflect different aspects and approaches that have been expored in Bielefeld during the last 10 to 12 years, especially in the area of human-robot interaction. A common backbone of different projects during this time has been a system-oriented research that is only possible in strongly collaborative environments. Re-appearing discussions have been the assessment of ground truth data, the re-producability of results, the question of cognitively adequate modelling, and the small size (resp. costs) of typical datasets. I will discuss these aspects with regard to a number of exemplary demonstration systems gaining insights and requirements for further research approaches.

Runtime Analysis of Robotics Software Architectures and Benchmarking

Sebastian Wrede (Bielefeld University)

The advent of robotics competitions, e.g., the Semantic Robot Vision Challenge at CVPR, the availability of publicly available datasets for specific robotics tasks, and dedicated efforts to raise the awareness for good experimental methodology, cf. the EURON GEM SIG, contributed significantly to an improved evaluation of cognitive robotics systems. However, most of the current approaches focus merely on aspects of the externally visible behaviour of these systems under certain task conditions. From a software engineering perspective, this does not yield many insights on the

internal structures and dynamics of the robotics systems under test. Given that scientific advancements on architectural organization and system integration techniques are key issues for managing the complexity of future cognitive systems, the talk claims that future benchmarks need to explicitly take into account the level of the system architecture to gain additional insights from system evaluation. This talk will highlight the importance of architectural support for runtime analysis of robotics architectures as a first step towards a more informative benchmarking of cognitive robots from a software engineering perspective.

Benchmarking of Motor Primitives Learning

Jochen Steil (Bielefeld University)

We report on work-in-progress from the EU-FP7 IP AMARSi: Adaptive Modular Architectures for Rich Motor Skills. The project comprises a number of experts for learning of Movement Primitives that use a diversity of dynamical systems approaches including DMP and reservoir networks.

For internal benchmarking and comparison of methods, AMARSi develops a MATLAB-based framework for automatic evaluation of such learning methods.

The AMARSi benchmark V1 is composed of a movement generation task in the 2D plane using for training handwriting motions provided by the EPFL - LASA (Aude Billard). The benchmark focuses on interpolation and generalization abilities and particularily on robustness against perturbation and ability to avoid obstacles.

The AMARSi benchmark V2 includes also learning of inverse kinematics and its generalization. In addition to measuring the accuracy of the static inverse kinematic mapping, a feedback-control scenario in which we use the ”webots” simulation of the humanoid robot iCub is under development. In the mid-term, AMARSi plans to make these benchmarking frameworks available to the scientific community.

The Turing Test Light - a behavioral way to evaluated interactive systems

Jan de Ruiter (Bielefeld University)

In evaluating interactive artificial systems, it is often not trivial to assess how good a particular implementation actually is. Often, semi-experimental designs with user questionnaires are used, but this method has well-documented disadvantages. Nevertheless, the assessment of the quality of implementations of interactive behavior by cognitive systems is important, not only to establish not how successful one's implementation is in an absolute sense, but also to measure relative progress within a project.

In this short contribution, I will present the idea of using the well-known "Wizard of Oz" situation (which is normally used to have human intelligence fill in the missing parts in an artificial system) as a "light" version of the famous Turing test for Artificial Intelligence.

I will show that by combining the "Turing Test Light" with signal detection theory and speed/accuracy trade-off analysis, this method could be a valuable additional evaluation method that avoids some of the pitfalls of the existing approaches.