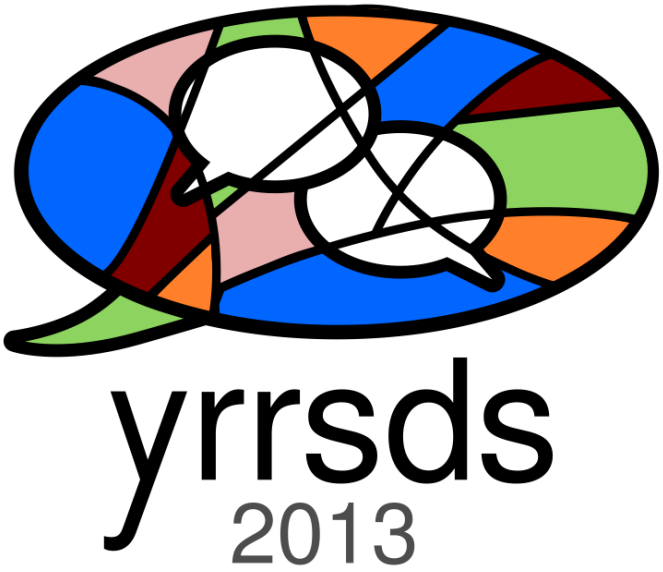
PROCEEDINGS OF THE 9TH ANNUAL

YOUNG RESEARCHER'S ROUNDTABLE ON SPOKEN DIALOGUE SYSTEMS



Supèlec, Metz, France
20^{th} - 21^{st} August, 2013
http://yrrsds.org

SPONSORS

Microsoft®
Research

ENDORSEMENTS

The Association for Computational Linguistics

The Special Interest Group on Discourse and Dialogue

International Speech Communication Association

9TH ANNUAL YOUNG RESEARCHER'S ROUNDTABLE ON SPOKEN DIALOGUE SYSTEMS (YRRSDS)

ORGANISERS

Layla El Asri
Orange Labs / Supèlec Metz, France
layla.elasri@orange.com

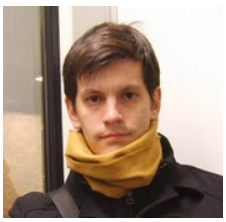
Nina Dethlefs
Heriot-Watt University, Edinburgh, UK
n.s.dethlefs@hw.ac.uk

Matthew Henderson
Cambridge University, UK
matthen@gmail.com

Casey Kennington
University of Bielefeld, Germany
ckennington@cit-ec.uni-bielefeld.de

Chris Mitchell
North Carolina State University, USA
cmmitch2@ncsu.edu

Niels Schütte
Dublin Institute of Technology, Ireland
niels.schuette@gmail.com



Martín Villalba
University of Potsdam, Germany
villalba.martin@gmail.com

Denis Baheux
Supèlec Metz, France
Denis.Baheux@supelec.fr

ADVISORY COMMITTEE

Mohammad Ahadi,
Amirkabir University of Technology

Srinivas Bangalore,
AT&T Research

Luciana Benotti,
Universidad Nacional de Còrdoba

Dan Bohus,
Microsoft Research

Rolf Carlson,
KTH Royal Institute of Technology

Maxine Eskenazi,
Carnegie Mellon University

Kallirroi Georgila,
University of Southern California

Gary Geunbae Lee,
Pohang University of Science and Technology

Jim Glass,
Massachusetts Institute of Technology

Kristiina Jokinen,
University of Helsinki

Tatsuya Kawahara,
Kyoto University

John Kelleher,
Dublin Institute of Technology

Alex Lascarides,
University of Edinburgh

Oliver Lemon,
Heriot-Watt University

Diane Litman,
University of Pittsburgh

Salam Mahbubush Khan,
Alabama A&M University

Wolfgang Minker,
University of Ulm

Sebastian Möller,
Technische Universität Berlin

Mikio Nakano,
Honda Research Institute

Carolyn Penstein Rosé,
Carnegie Mellon University

David Schlangen,
Universität Bielefeld

Stephanie Seneff,
Massachusetts Institute of Technology

Gabriel Skantze,
KTH Royal Institute of Technology

Amanda Stent,
AT&T Research Labs

David Traum,
University of Southern California

Marilyn Walker,
University of California Santa Cruz

Nigel Ward,
University of El Paso

Jason Williams,
Microsoft Research

Steve Young,
University of Cambridge

TABLE OF CONTENTS

PROGRAMME

Tuesday, 20 August, 2013
8:30 registration, poster set up
9:00 welcome
9:30 **first roundtable discussion**
11:00 talk by **Olivier Pietquin**
11:30 talk by **Kristiina Jokinen**
12:00 lunch (provided)
13:30 **poster sessions**
15:30 coffee
16:00 **second roundtable discussion**
17:30 end of day 1
18:30 dinner

Wednesday, 21 August, 2013
9:00 **special session on tools, toolkits, and demos**
10:30 coffee break
10:50 talk by **Casey Kennington (GET)**
11:00 talk by **Amanda Stent (AT&T)**
11:30 talk by **Jason Williams (MSR)**
12:00 lunch (provided, extended so we can return to posters)
14:00 **third roundtable discussion**
15:30 coffee break
16:00 **industry / academic panel**
17:30 conclusion
18:00 end of YRRSDS 2013

PREFACE

We are delighted to welcome you to the Ninth Young Researchers' Roundtable on Spoken Dialogue Systems in Metz, France. YRRSDS is a yearly event that began in 2005 in Lisbon, followed by Pittsburgh, Antwerp, Columbus, London, Tokyo, Portland, and Seoul.

The aim of the workshop is to promote the networking of students, post docs, and junior researchers working in research related to spoken dialogue systems in academia and industry. The workshop provides an open forum where participants can discuss their research interests, current work, and future plans.

This year we have 20 registered participants representing many different countries. The roundtable will also feature 7 guest participants who will give presentations representing their work, and a panel discussion. Of those, we have invited representatives from our sponsors, Amanda Stent (AT&T) and Jason Williams (Microsoft Research). Our invited academic speakers are Kristiina Jokinen (Univeristy of Helsinki) and Olivier Pietquin (Supèlec Metz). Added to these, we have three more participants on the academic / industry panel: David DeVault (ICT, University of Southern California), Deepak Ramachandran (Nuance), and Graham Wilcock (University of Helsinki).

We are grateful for the continued sponsorship from AT&T and Microsoft Research, both who have been very supportive in years past. We are also grateful to Global Educational Technologies for sponsoring YRRSDS this year. We are endorsed by ACL, SIGdial, and ISCA, and are delighted and grateful for their support. We extend a special thanks to our local organizers, Layla El Asri and Denis Baheux who organized the venue, accomodations, meals, and the social event.

Finally, we thank you the participants of YRRSDS. You are what make it work. We are excited for this year's group and look forward to interesting discussion!

Layla El Asri
Nina Dethlefs
Matthew Henderson
Casey Kennington
Chris Mitchell
Niels Schütte
Martín Villalba
Denis Baheux

(YRRSDS-2013 organisers)

INVITED GUESTS

SPONSOR TALK

AMANDA STENT

AT&T

Amanda Stent
AT&T
http://www.amandastent.com/

Biographical Sketch Dr. Amanda Stent works on spoken dialog, natural language generation and assistive technology. She is currently a Principal Member of Technical Staff at AT&T Labs - Research in Florham Park, NJ and was previously an associate professor in the Computer Science Department at Stony Brook University in Stony Brook, NY. She holds a PhD in computer science from the University of Rochester. She has authored over 60 papers on natural language processing and is one of the rotating editors of the journal Dialogue and Discourse. She was recently elected the President of SIGdial.

SPONSOR TALK

JASON WILLIAMS

Microsoft Research

Jason Williams
Microsoft Research
http://research.microsoft.com/en-us/people/jawillia/

Biographical Sketch Dr. Jason Williams is a researcher at Microsoft Research. Before that and since 2006, he was a Principal Member of Technical Staff at AT&T Labs Research. He received a BSE in Electrical Engineering from Princeton University in 1998, and at Cambridge University he received an M Phil in Computer Speech and Language Processing in 1999 and a Ph.D. in Information Engineering in 2006. His main research interests are dialogue management, the design of spoken language systems, and planning under uncertainty. He is currently Editor-in-chief of the IEEE Speech and Language Processing Technical Committee's Newsletter. He is also on the Science Advisory Committee of SIGdial, currently holds the position of vice president of SIGdial, and the board of directors of AVIxD. Prior to entering research, he built commercial spoken dialogue systems for Tellme Networks (now Microsoft), and others.

ACADEMIC TALK

OLIVIER PIETQUIN

Supèlec Metz

Olivier Pietquin
Supèlec Metz, France
http://www.metz.supelec.fr/metz/personnel/pietquin/

Biographical Sketch: Olivier Pietquin obtained an Electrical Engineering degree from the Faculty of Engineering, Mons (FPMs, Belgium) in June 1999 and a PhD degree in April 2004. In 2011, he received the Habilitation Diriger des Recherches (French Tenure) from the University Paul Sabatier (Toulouse, France). He joined the FPMs Signal Processing department (TCTS Lab.) in September 1999. In 2001, he has been a visiting researcher at the Speech and Hearing lab of the University of Sheffield (UK). Between 2004 and 2005, he was a Marie-Curie Fellow at the Philips Research lab in Aachen (Germany). Now he is a Professor at the Metz campus of the Ecole Superieure d'Electricite (Supèlec, France), and headed the "Information, Multimodality & Signal" (IMS) research group from 2006 to 2010 when the group joined the UMI 2958 (GeorgiaTech - CNRS). In 2012, he heads the Machine Learning and Interactive Systems group (MaLIS). From 2007 to 2011, he was also a member of the IADI INSERM research team (in biomedical signal processing). He is a full member of the UMI 2958 (GeorgiaTech - CNRS) since 2010 and coordinates the computer science department of this joint lab. Since 2010, Olivier Pietquin sits at the IEEE Speech and Language Technical Committee and he is a Senior IEEE member since 2011. His research interests include spoken dialog systems evaluation, simulation and automatic optimisation, machine learning, speech and signal processing. He authored or co-authored more than 100 publications in these domains.

ACADEMIC TALK

KRISTIINA JOKINEN

University of Helsinky

Kristiina Jokinen
University of Helsinky, Finland
http://www.ling.helsinki.fi/ kjokinen/

Biographical Sketch: Dr. Kristiina Jokinen did her first degree in mathematics at the University of Helsinki and her Ph.D in computational linguistics at the University of Manchester Institute of Science and Technology, where her thesis involved development of the Constructive Dialogue Model. She was a fellow at the Nara Institute of Science and Technology, and later an invited researcher at ATR Interpreting Telecommunications Labratory in Kyoto, Japan. She set up the research lab CELE (Centre for Evolutionary Language Engineering) in Belgium. She is currently an adjunct professor of language technology at the University of Helsinki, where she teaches and researches in the areas of dialogue management, non-verbal communication and multimodality and is the project director of the research group 3I (Intelligent Interactive Information Systems). She is also a visiting professor at the University of Tartu and an adjunct professor at the University of Tampere. She has been involved in SIGdial for many years and is currently the secretary/treasurer.

PANELIST

DAVID DEVAULT

Institute for Creative Technologies, University of Southern California

David DeVault
Institute of Creative Technologies, University of Southern California
http://people.ict.usc.edu/ devault/

Biographical Sketch: Dr. David DeVault is currently a research assistant professor in the Department of Computer Science at the University of Southern California, and a research scientist at the Institute for Creative Technologies. My research is in natural language dialogue systems, which converse and interact with human speakers in a natural language such as English. The central theme is to enable dialogue systems to respond to the inevitable uncertainties of communication in a way that is more flexible, more robust, and more human-like. He received is BS in Engineering and Applied Science at Caltech in 2000, then went on to Rutgers where he earned a M.A. Philosophy and Ph.D in Computer Science in 2004 and 2008, respectively. In 2008, he was a postdoctoral research associate at ICT. Starting in 2010 he became a research scientist at ICT, and in 2011 became a research assistant professor at the USC department Computer Science.

PANELIST

DEEPAK RAMACHANDRAN

Research Scientist, Nuance Communications, Ltd.

Deepak Ramachandran
Nuance Communications, Ltd.

Biographical Sketch: Deepak Ramachandran's background is in AI and machine learning. He did his PhD at the University of Illinois at Urbana-Champaign, working on the problem of Inverse Reinforcement Learning. He has worked as a Scientist at the Honda Research Institute, applying machine learning to problems in hybrid vehicle design, human-robot interaction and dialog systems. Recently he joined the Nuance Communications NLU lab in Sunnyvale, CA where he works on dialog management and advanced conversational prototypes.

PANELIST

GRAHAM WILCOCK

University of Helsinki, Finland

Graham Wilcock
University of Helsinki, Finland
http://www.ling.helsinki.fi/ gwilcock/

Biographical Sketch: Dr. Graham Wilcock is currently a university lecturer and adjunct professor at the University of Helsinki in Finland, where he has been since 2001. Prior to that, he spent many years in language technology, starting with Japanese-English machine translation at the University of Sheffield. He spent a number of years as a senior research scientist at Sharp Corporation working in machine translation, while simultaneously doing research in computational linguistics at UMIST in Manchester, where he completed his Ph.D. He has taught at various universities in the United Kingdom including the University of Manchester, the Open University, University of Hull Centre for Internet Computing, and Imperial College London. He has written at least one book, various book chapters as well as numerous other publications in syntactic parsing (notably for HPSG), language generation, human-robot interaction, as well as other topics.

POSITION PAPERS

List of Participants

Timo Baumann, Hamburg University, Germany

Crystal Chao, Georgia Institute of Technology, U.S.A.

Layla El Asri, Orange Labs / Supélec Metz, France

Tatiana Gasanova, Ulm University, Germany

Nadine Glas, Télécom-ParisTech, France

Matthew Henderson, University of Cambridge, U.K.

Casey Kennington, Bielefeld University, Germany

Hatim Khouzaimi, Orange Labs, Avingnon, France

Alejandra Lorenzo, Université de Lorraine, France

Yoichi Matsuyama, Waseda University, Japan

Raveesh Meena, KTH Royal Institute of Technology, Sweden

Christopher Mitchell, North Carolina State University, U.S.A.

Aasish Pappu, Carnegie Mellon University, U.S.A.

JonhHo Shin, Korea Telecom, South Korea

Dirk Schnelle-Walka, Technical University Darmstadt, Germany

Niels Schütte, Dublin Institute of Technology, Ireland

Stefan Ultes, Ulm University, Germany

Martin Villalba, University of Potsdam, Germany

Timo Baumann

Universität Hamburg
Natural Language Systems Division
Department of Informatics

baumann@informatik.uni-hamburg.de
http://www.timobaumann.de/work/

1 Research Interests

My research is geared towards **interaction management** in spoken dialogue. Specifically, I am interested in the **fine-grained timing** of dialogue and interaction-related phenomena. For a dialogue system to achieve the level of control that I think is necessary for good dialog behaviour, it is necessary for the system to run **incrementally**, that is, to process the user's utterance while it is ongoing, and to come up with partial conclusions about what the user is saying, what the system should answer and how certain this is. I am also interested in **proactively hypothesizing** about the near future, generating output that **predicts** a short distance into the future in order to overcome delays or to –gasp– cut short the user. While traditionally a dialogue system could only be sluggish or fast enough, a proactive system's timing must try to temporally align to the user (or to deliberately break the alignment). Of course, a system that acts incrementally and predictively should be able to also *act incrementally* itself, being able to change or extend plans that are already being executed as the situation unfolds. I believe that **prosody** plays a vital role in everyday conversation and that it is still too often ignored due to a prevalence of written language and a turn-taking paradigm based on ping-pong-style interaction. I believe that a leap in spoken dialogue systems design and performance will result from considering more fine-grained timing and prosodic information across the board.

1.1 Incremental Processing and Evaluation

In a modular system, an *incremental module* is one that generates (partial) output while input is still ongoing. I have thoroughly investigated the evaluation of such incremental processors (Baumann et al., 2011). The metrics we developed deal with how often hypotheses change (every change means that consuming modules have to re-process their input) and describe timing properties of events relative to their ideal detection. In incremental processing, there is a trade-off between the timing, the quality, and the stability of hypotheses: The earlier we hypothesize, the more likely the hypothesis is wrong, and the more often we may have to revise before arriving at a correct result.

I showed this influence for incremental speech recognition (iSR) derived a measure of certainty from the different timing measures and also devised algorithms that improve these incremental properties for iSR using generic filtering mechanisms (Baumann et al., 2009). As part of our venture into incremental analysis, we built a toolkit to process and visualize incremental data (Malsburg et al., 2009), and the incremental processing toolkit INPROTK (Baumann and Schlangen, 2012b).

1.2 Predictive Processing

In an SDS, some processing latencies are inevitable. Hence, for reactions to be *right on time*, they must be issued *before the fact*. In other words, for natural interaction, an SDS must anticipate future events (e. g. that a back-channel or speaker change will be required soon) and predict when exactly to react. I am particularly interested in the micro-timing of these predictions, and built a system that synchronoulsy completes words (and full turns) while the speaker is still speaking them (Baumann and Schlangen, 2011), showing that end-to-end incremental processing is possible in real time. I believe that good system timing no longer means "as quickly as possible" but that precise timing will become possible and important.

1.3 Incremental Speech Synthesis

Recently, I have worked on incremental, *just-in-time* speech synthesis (iSS), showing that a system can start speaking with very little utterance-initial processing (Baumann and Schlangen, 2012a) which leads to better system response times and allows for more natural behaviour (Buschmeier et al., 2012). In our approach, synthesis is tightly integrated into the SDS data structures, allowing for seamless, immediate, and on-the-fly adaptation of system utterances. Our experiments show that the interactive strategies enabled by iSS are preferred in interactive situations (Baumann and Schlangen, 2013b), and that interactional adequacy is an important aspect when judging synthesis systems (Baumann and Schlangen, 2013a).

1.4 Future Work

Having just completed my PhD (Baumann, 2013), I hope to be able to further improve the 'conversational' abilities of spoken dialogue systems by advancing the speech-related capabilities with respect to dialogue specifics, and by exploring the real-time integration of additional sources of information into the system.

2 Future of Spoken Dialogue Research

I believe that in the future, dialogue systems will appear as **conversational assistants** in many areas, such as hospitals, for elderly people, in tutoring (not only for foreign language learning, but in all areas), and one of the natural interfaces of general-purpose life-long digital assistants.

Such a digital assistant will likely appear in multiple modalities. Often, blending multiple modalities will be the method of choice, calling for a thought-out way of integrating speech into the multi-modal system.

While human-like behaviour is not needed or could even distract in simple task-oriented systems, human-like behaviour may be more important for future applications, as they will be less recognized as tools but as real interlocutors. For better intuitivity, **interaction behaviour** (turn-taking, and -yielding, understanding and hinting below the content level) must be improved.

3 Suggestions for Discussion

VUI or SDS? Apple's Siri has shown the tremendous success that a well-designed speech application can have. However, Siri is 'just' a VUI rather than a full SDS and far from being a conversational agent. However, airplanes only ever took off when engineers stopped trying to flap their wings. How much naturalness will be required for future SDSs? Is naturalness really the key to successfull dialog applications?

Turn-by-turn *vs.* continuous interaction: Engineers of applied dialogue systems think of "barge-ins" when they talk about flexibility in their system's turn-taking scheme. While the *turn-by-turn paradigm* helps to arrange contributions to dialogue conceptually, I believe that it is becoming a handicap in dialogue research and development, as it barely reflects "real" dialogue, in which people constantly interact, give feedback about understanding, consent, etc. with much of this interaction happening on the sub-word level.

Architectures for Highly Interactive Dialogue Systems To manage conversational dialogue, systems must manage uncertainty, incrementality, frequent revision of plans, and deep linguistic and knowledge representations, all at the same time. Furthermore, all these aspects should be integrated to some degree, as they are by the human mind. How shall we model the coupling of components, interfaces to internal knowledge of components, and interaction on different time-scales, preferably in a structured, learnable, statistical model?

References

Timo Baumann and David Schlangen. 2011. Predicting the Micro-Timing of User Input for an Incremental Spoken Dialogue System that Completes a User's Ongoing Turn. In *Procs of SigDial 2011*, Portland, USA.

Timo Baumann and David Schlangen. 2012a. INPRO_iSS: A component for just-in-time incremental speech synthesis. In *Procs. of ACL System Demos*, Jeju, Korea.

Timo Baumann and David Schlangen. 2012b. The INPROTK 2012 release. In *Proceedings of SDCTD*, Montréal, Canada.

Timo Baumann and David Schlangen. 2013a. Interactional adequacy as a factor in the perception of synthesized speech. In *Proceedings of Speech Synthesis Workshop (SSW8)*. to appear.

Timo Baumann and David Schlangen. 2013b. Open-ended, extensible system utterances are preferred, even if they require filled pauses. In *Proceedings of SigDIAL*. to appear.

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Procs. of NAACL-HLT 2009*, pages 380–388, Boulder, USA.

Timo Baumann, Okko Buß, and David Schlangen. 2011. Evaluation and optimisation of incremental processors. *Dialogue & Discourse*, 2(1):113–141. Special Issue on Incremental Processing in Dialogue.

Timo Baumann. 2013. *Incremental Spoken Dialogue Processing: Architecture and Lower-level Components*. Ph.D. thesis, Universität Bielefeld, Germany.

Hendrik Buschmeier, Timo Baumann, Benjamin Dorsch, Stefan Kopp, and David Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Procs. of SigDial*, Seoul, Korea.

Titus von der Malsburg, Timo Baumann, and David Schlangen. 2009. TELIDA: A Package for Manipulation and Visualisation of Timed Linguistic Data. In *Procs. of SigDial 2009*, pages 302–305, London, UK.

Biographical Sketch

Timo Baumann is a researcher and instructor at Universität Hamburg, Germany, with Wolfgang Menzel. He recently completed his PhD on incremental spoken dialogue processing at Bielefeld University under the supervision of David Schlangen.

Timo studied computer science, phonetics and linguistics in Hamburg, Geneva, and Granada and received his master's degree in 2007 for work on prosody analysis at IBM Research. He previously worked at the Universities of Potsdam, Bielefeld, and Stockholm before returning to Hamburg last year.

In his free time, Timo likes to go hiking or cycling and sings in a choir. He prefers organic food and is interested in renewable energies.

Crystal Chao

School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia 30332, USA

cchao@gatech.edu
cc.gatech.edu/social-machines/cchao

1 Research Interests

My research focuses on controlling the timing of **turn-taking** within a **situated dialogue** to achieve fluent **human-robot interaction (HRI)**. I develop autonomous **multimodal** control for the robot's **social behavior**, which includes speech, gesture, and eye gaze. In addition, I am interested in how **semantics** and **pragmatics** should be integrated within a robot's cognition and action control in order to create natural and generalizable dialogue-based interactions.

1.1 Balance of interaction control

Turn-taking is a process of managing shared resources in an interaction. In a spoken dialogue, the fundamental shared resource is the speaking floor; in a situated dialogue, other such resources may exist, such as shared physical space. The agent who owns the shared resource at a given time has control over the interaction outcome. Thus, appropriate turn-taking is an expression of appropriate balance of control, which is necessary for fulfilling the interaction goals.

Balance of control over shared resources has effects on the outcome of a joint activity. In a previous study I conducted, participants collaborated with a robot on a manipulation-based puzzle, the Towers of Hanoi, and took more initiative in the task when the robot interrupted its speech and manipulation actions than when it did not (Chao and Thomaz, 2012). I also found in previous work that human teachers lost track of their mental models of a teaching task when the robot asked questions too often (Cakmak et al., 2010). Thus, it's important for an autonomous system to exert control over resources in an intentional and appropriate way for a given situation.

1.2 Parametrized floor regulation

People adapt their own turn-taking styles to different social situations. For example, they take shorter turns and interrupt themselves more often in the presence of someone with more authority. The style communicates knowledge of one's social role as well as functionally regulates the priority of control over shared resources.

To achieve this capability in robots, I developed a system called CADENCE: Control Architecture for the Dynamics of Embodied Natural Coordination and Engagement (Chao and Thomaz, 2013). CADENCE supports the setting of turn-taking parameters such as relative floor time, response delay, interrupts, and overlap and silence tolerance. Human users attributed different personalities to the system when they interacted with different settings, and also adjusted their speaking behavior in response.

1.3 Modeling information flow

An important factor in the timing of turn-taking is the flow of information between interaction participants. I previously ran a Wizard of Oz study in which a robot played the imitation game "Simon says" with human participants (Chao et al., 2011). The most robust indicator of human response delay to robot turns was the time at which the robot had communicated enough information to respond to in a semantically appropriate way (that is, discounting barge-ins or simultaneous starts), across modalities (speech and gesture). We defined this as the point of *minimum necessary information (MNI)*.

The principle of MNI describes the importance of alignment between semantic understanding and behavior generation in communicative acts. The system can choose to interrupt itself for efficiency's sake when the MNI point has passed (Thomaz and Chao, 2011). User turns prior to the MNI point should not be considered as responses to the system's current turn. Thus, systems that support incremental understanding and generation should lead to improved turn-taking timing.

1.4 Timed Petri nets (TPNs) for turn-taking

Achieving appropriate turn-taking timing on a social robot requires the control of multiple modalities, such as gaze, speech, and gesture. Each of these modalities is unique in how it represents bottlenecking resources, encodes semantic information to be communicated, interprets perceptual data from a human, generates temporally extended behavioral actions, and interacts with other instrumental actions on the robot such as manipulation or attentional processes.

The representation I have been using for controlling autonomous turn-taking interactions is the timed Petri net (TPN). Cross-modality interactions are cumbersome to represent as a single finite state machine or multiple interacting finite state automata. Since TPNs are

meant for representing workflows featuring synchronization, concurrency, and mutual exclusions, I have found them to be a better fit for expressing the behavioral processes involved in real-time turn-taking. They offer a good abstraction level and shared representation for modeling, control, and analysis through simulation (Chao and Thomaz, 2012).

2 Future of Spoken Dialog Research

Advances in perception will create the first wave of ubiquitous dialogue systems. Currently, one-off voice commands are highly viable. People can use speech to trigger simple actions with Apple Siri or a living room Kinect, offering a convenient hands-free interface. With increased awareness of a user's ongoing state or activity in the home or in a car, a more sophisticated ongoing dialogue becomes possible. In addition to commanding device state (e.g. for lights, a thermostat, a television, speakers, etc.), a user can request information relevant to a current task such as recipe steps or entertainment recommendations, as well as refine parameters over time. With enough data about a user's habits and usage, the system can offer suggestions and report anomalies (e.g. "Did you want the garage door closed at this hour?").

Eventually, mobile manipulators will become more capable and affordable and start to appear in homes. The physical aspect of the robot will extend the impact of automated action possible, but there will also be higher risk associated with error. The robot's embodiment will introduce the issue of being physically intrusive and will also elicit different behavior from the human user.

An issue for any such system is knowing when it is actually appropriate to take an action. For example, a user may want to talk about a system action without wanting to trigger it. It can also be ambiguous when the system should take initiative; a system-initiated event could be considered necessary to daily functioning or disruptive, be it an alarm, a calendar reminder, or a preemptive beer from the fridge. System confirmations could be considered a welcome buffer against error or an annoyance. This opinion could change depending on the user's activity state, such as having guests over, talking on the phone, showering, or watching television.

Another challenge for these future dialogue systems will be the representation of semantics. Especially for a robot, real-world actions and recognition processes must align with spoken communication in order to be interpreted intuitively by a human user. Thus, perception, action, and language need to be tightly coupled in these interactive systems. Bu we will also want systems with differing embodiments to access shared databases describing grounded knowledge and how to talk about it. These semantic representations will need to be expressive, easy to design or learn, and transferable between systems.

3 Suggestions for Discussion

- Statistical vs. knowledge-based modeling: What aspects of dialogue are best suited for each approach, in terms of transfer, performance, and ease of development? When should we design, and when should we collect more data?
- Universal domain specification: How can we easily swap domains between systems of contrasting capabilities? How can we share or crowdsource efforts in domain authoring in order to reduce development time of dialogue experts spent on it?
- System abstraction levels: Can we expose or redraw standard black boxes in order to create new capabilities? For example, can discourse context influence acoustic-level recognition parameters?

References

Maya Cakmak, Crystal Chao, and Andrea L. Thomaz. 2010. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2(2):108–118.

Crystal Chao, Jinhan Lee, Momotaz Begum, and Andrea L. Thomaz. 2011. Simon plays Simon says: The timing of turn-taking in an imitation game. *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 235–240.

Crystal Chao and Andrea L. Thomaz. 2012. Timing in multimodal turn-taking interactions: Control and analysis using timed Petri nets. *Journal of Human-Robot Interaction*, 1(1):4–25.

Crystal Chao and Andrea L. Thomaz. 2013. Controlling social dynamics with a parametrized model of floor regulation. *Journal of Human-Robot Interaction*, 2(1):4–29.

Andrea L. Thomaz and Crystal Chao. 2011. Turn-taking based on information flow for fluent human-robot interaction. *AI Magazine: Special Issue on Dialog with Robots*, 32(4):53–63.

Biographical Sketch

Crystal Chao is a Ph.D. candidate in Robotics and President's Fellowship recipient at the Georgia Institute of Technology, where she is member of the Socially Intelligent Machines Lab directed by Dr. Andrea L. Thomaz. Her research focuses on human-robot turn-taking and socially interactive learning. She received a B.S. in Computer Science from the Massachusetts Institute of Technology in 2008.

Layla El Asri

Orange Labs / UMI 2958 (IMS-MaLIS, Supélec Metz)
38-40 rue du Général Leclerc 92130 Issy-les-Moulineaux, France
2 rue Édouard Belin 57070 Metz, France

layla.elasri@orange.com

1 Research Interests

Hand-crafting the behaviour of the Dialogue Manager (DM) of a Spoken Dialogue System (SDS) is cumbersome and might result in a DM difficult to transfer to other tasks. For this reason, Reinforcement Learning (RL) has been extensively used to model dialogue management and automatically learn an optimal behaviour (Singh et al., 1999). Nevertheless, many parameters of the RL framework are still hand-crafted such as the state space or the reward function. Paek (2006) even qualified the reward function as "most hand-crafted aspect of reinforcement learning". My Ph.D. aims to propose a methodology for the inference of a reward function from data instead of designing it according to experience and intuition about the given task. The problem we are trying to solve is given in Definition 1.

Definition 1 (Reward inference problem) *Infer a reward function from a corpus of N dialogues* $\mathcal{D} = (D_i)_{i \in 1..N}$ *manually evaluated with a numerical performance score* $P_i \in \mathbb{R}$.

1.1 Past and current work

We have first proposed two algorithms solving the problem in Definition 1 by computing diffuse reward functions: a diffuse function gives a reward after each transition between two states of a dialogue (El-Asri et al., 2012). They thus highly depend on the quality of the state space which must take into account the many parameters coming into play in dialogue evaluation (quality of speech recognition, system cooperativity,...).

On a simple example, where dialogue evaluation was quite accurately described by the state space, we have shown that using one of our diffuse functions resulted in a faster learning then when a sparse function, giving a reward only at the end of each dialogue, was used (El-Asri et al., 2013). The dialogues and their evaluations were simulated with a light version of the TownInfo system (Lemon et al., 2006). The SDS provides information about local restaurants given a price range, an area and a type of food.

An important obstacle in automatic reward computation is the non-observability of task completion and speech recognition failures. We are working on a method which uses $\mathcal{D}$ to overcome online non-observability of the task completion.

Another one of our current challenges is to also infer an optimised state space from $\mathcal{D}$. We have designed an RL-based appointment scheduling system and had it evaluated by 383 users. The resulting corpus of 1734 dialogues was named DINASTI (DIalogues with a Negotiating Appointment SeTting Interface). The system was conceived in accordance with Grice's cooperativity principles (Grice, 1989; Bernsen et al., 1996; Dybkjaer et al., 1996). To be fully cooperative, a system has to adapt its behaviour to user expertise and be able to efficiently track dialogue evolution. Therefore, we have used RL to provide the system with these capacities. In concrete terms, a specific set of actions was embedded in the SDS like choosing which dialogue and task initiative strategies, the amount of information to provide, when to ask for a confirmation, *etc*. Our goal is now to infer from this corpus a state space which adequately deals with user expertise, speech recognition problems, *etc*. using automatically computable dialogue features and techniques of RL in continuous spaces.

1.2 Future work

We plan to work on a better management of speech recognition performance by the state space inferred from $\mathcal{D}$. Indeed, it takes an important place in users evaluation (Walker et al., 1998; Larsen, 2003) and must thus be handled correctly by our diffuse reward functions. Another feature that is desirable is active learning: since it is usually not easy to collect many evaluated dialogues, we aim to provide SDS with the possibility to act online to reduce the uncertainties about their learning.

2 Future of Spoken Dialog Research

- Dialogue systems integrated into smart environments (cars, homes). This requires to handle multiple user commands concerning many different tasks (dictating an email, managing some device, searching for an information). This level of complexity is a challenge for natural language understanding and

dialogue management.

- RL-based dialogue management should be brought within the reach of developers who are not experts in machine learning. Corpora analysis should enable to automatically extract RL parameters such as the state space or the reward function.
- Towards a more natural dialogue. Incremental systems allowing a different dialogue logic than the turn taking one, systems that can handle a broader range of utterances seem to be the next gap for spoken dialogue technology.

3 Suggestions for Discussion

- Automatic conception of SDS: what still needs to be done ? Can we extend it to RL-based SDS ? What are the main challenges ? How can we combine corpora analysis and best practices to design ready-to-use SDS ?
- Active learning for RL-based SDS: how can we integrate active learning, what degrees of freedom should be left to the system ? How can an SDS use active learning to track possible trouble in dialogue management ?
- Inspiration from the philosophy of language: theories from the philosophy of language have been adapted to SDS (Searle's speech act, Grice's cooperativity principles). How do the two disciplines interact ? How can the philosophy of language help us conceive even more efficient SDS ? Should we turn to more human-like dialogue or to SDS specialised in the achievement of a given task ?

References

Niels Ole Bernsen, Hans Dybkjr, and Laila Dybkjr. 1996. Principles for the design of cooperative spoken human-machine dialogue. In *Proceedings of ICSLP*, pages 729–732.

Laila Dybkjaer, Niels Ole Bernsen, and Hans Dybkjaer. 1996. Grice incorporated: Cooperativity in spoken dialogue. In *Proceedings of COLING*.

Layla El-Asri, Romain Laroche, and Olivier Pietquin. 2012. Reward function learning for dialogue management. In *Proceedings of STAIRS*.

Layla El-Asri, Romain Laroche, and Olivier Pietquin. 2013. Reward shaping for statistical optimisation of dialogue management. In *Proceedings of SLSP (to be published)*.

Paul Grice, 1989. *Studies in the Way of Words*, chapter Logic and Conversation. Harvard University Press, Cambridge MA.

Lars Bo Larsen. 2003. Issues in the evaluation of spoken dialogue systems using objective and subjective measures. In *Proceedings of IEEE ASRU*, pages 209–214.

Oliver Lemon, Kallirroi Georgila, James Henderson, and Matthew Stuttle. 2006. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: Generic slot-filling in the talk in-car system. In *Proceedings of EACL*.

Tim Paek. 2006. Reinforcement learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment. In *Proceedings of Interspeech, Dialog-on-Dialog Workshop*.

Satinder Singh, Michael Kearns, Diane Litman, and Marilyn Walker. 1999. Reinforcement learning for spoken dialogue systems. In *Proceedings of NIPS*.

Marilyn A. Walker, Jeanne C. Fromer, and Shrikanth Narayanan. 1998. Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email. In *Proceedings of COLING/ACL*, pages 1345–1352.

Biographical Sketch

Layla El Asri graduated from Grenoble INP Ensimag, France in 2011. She has an engineering degree in computer science and a research master in artificial intelligence. She has joined Orange Labs in Issy-les-Moulineaux, France for an internship in 2011 and she stayed as Ph.D. student with the collaboration of IMS-MaLIS in Supélec Metz (UMI 2958). Her Ph.D. supervisors are Olivier Pietquin in IMS-MaLIS and Romain Laroche in Orange Labs.

Tatiana Gasanova

Dialogue Systems Research Group
Ulm University
Albert-Einstein-Allee 43
89081 Ulm, Germany

tatiana.gasanova@uni-ulm.de
nt.uni-ulm.de/gasanova

Research Interests

In the area of **Spoken Dialogue Systems** (SDS), my interest lies in the natural language processing and text categorization. I am also interested in **Machine Learning**, **Information Retrieval** and **Data Mining**. The general idea is to use evolutionary algorithms and other heuristics to develop efficient classification algorithms which are able to compete with the existing models.

1.1 Natural Language Call Routing

Natural language call routing can be treated as an instance of topic categorization of documents (where the collection of labeled documents is used for training and the problem is to classify the remaining set of unlabeled test documents) but it also has some differences. For instance, in document classification there are much more terms in one object than in single utterance from call routing task, where even one-word utterances are common.

The most similar work has been done by A. Albalate, D. Suendermann, R. Pieraccini, and W. Minker. They have worked on the data with the same structure: the focus was on the problem of big part of non-labeled data and only few labeled utterances for each class, methods of matching the obtained clusters and the given classes have also been considered; they provided the comparison of several classification methods that are able to perform on the large scale data.

The information retrieval approach for call routing is based on the training of the routing matrix, which is formed by statistics of appearances of words and phrases in a training set (usually after morphological and stop-word filtering). The new caller request is represented as a feature vector and is routed to the most similar destination vector. The most commonly used similarity criterion is the cosine similarity. The performance of systems, based on this approach, often depends on the quality of the destination vectors.

We propose a new term relevance estimation approach based on fuzzy rules relevance for fuzzy classifier (H. Ishibuchi, T. Nakashima, and T. Murata., 1999) to improve routing accuracy. We have also used a decision rule different from the cosine similarity. We assign relevancies to every destination (class), calculate the sums of relevancies of words from the current utterance and choose the destination with the highest sum.

The database for training and performance evaluation consists of about 300.000 user utterances recorded from caller interactions with commercial automated agents. The utterances were manually transcribed and classified into 20 classes (call reasons), such as *appointments*, *operator*, *bill*, *internet*, *phone* or *video*. Calls that cannot be routed certainly to one reason of the list are classified to class *_TE_NOMATCH*.

A significant part of the database (about 27%) consists of utterances from the "garbage" class (*_TE_NOMATCH*). Our proposed approach decomposes the routing task into two steps. On the first step we divide the "garbage" class into the set of subclasses by one of the clustering algorithms and on the second step we define the call reason considering the "garbage" subclasses as separate classes. We apply genetic algorithms with the whole numbers alphabet, vector quantization network and hierarchical agglomerative clustering in order to divide "garbage" class into subclasses. The reason to perform such a clustering is due to simplify the detection of the class with non-uniform structure.

Our approach uses the concept of salient phrases: for each call reason (class) only 300 words with the highest term relevancies are chosen. It allows us to eliminate the need for the stop and ignore word filtering.

1.2 Text Categorization

Nowadays, Internet and the World Wide Web generate a huge amount of textual information. It is increasingly important to develop methods of text processing such as text categorization. Text categorization can be considered to be a part of natural language understanding, where there is a set of predefined categories and the task is to automatically assign new documents to one of these categories. There are many approaches to the analysis and categorization of text, but they could be roughly divided into statistical approaches, rule-based approaches and their combinations. Furthermore, the

method of text preprocessing and text representation influences the results that we obtained even with the same methods.
Related work was done by the participants of the fourth edition DEFT text mining campaign (Bechet et al., 2008; Charnois at al., 2008; Charton et al., 2008; Cleuziou and Poudat, 2008; Plantie et al., 2008; Trinh et al., 2008), which have worked with the same data using some classic approaches for text categorization and combinations of standard algorithms with the original ideas.
The proposed approach consists of preprocessing step, when we extract all words from the train set regardless of the case of the letters and excluding the punctuation. Then using our formula for word relevance estimation and applying hierarchical clustering algorithm we obtain a set of clusters and assign a common value to the whole cluster. However these common values do not provide the maximum classification quality, and therefore we suggest using hybrid genetic algorithm to improve the values corresponding to a single category and coevolutionary genetic algorithm with cooperation scheme (Potter and De Jong, 2000) to improve all values in parallel.

Future of Spoken Dialog Research

In the next years, researchers will not only improve and solve a lot of issues in Spoken Dialogue Systems but these systems will be common around the world. People will interact with computers using speech and it will not be something unusual. There will be smart homes which will automate the daily routine of many people. It is very important to find out which kind of dialogue users prefer and SDS should be able to adapt to concrete user. We will also find the new implementation of SDS and the ways to integrate Spoken Dialogue Systems more into the real life.

Suggestions for discussion

- Natural language or key words: which is better in call routing systems?
- Which approach is better: rule-based or statistical?
- How should we handle objects which do not fit in predefined classes?

References

A. Albalate, D. Suendermann, R. Pieraccini, and W. Minker. 2009. *Mathematical Analysis of Evolution, Information, and Complexity*, Wiley, Hoboken, USA.

F. Bechet, M.E. Beze and J.-M. Torres-Moreno. 2008. *Proceedings of the 4th DEFT Workshop* (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, 27-36.

T. Charnois, A. Doucet, Y. Mathet and F. Rioult. 2008. *Proceedings of the 4th DEFT Workshop* (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, 37-46.

E. Charton, N. Camelin, R. Acuna-Agost, P. Gotab, R. Lavalley, R. Kessler and S. Fernandez. 2008. *Proceedings of the 4th DEFT Workshop* (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, 47-56.

G. Cleuziou and C. Poudat. 2008. *Proceedings of the 4th DEFT Workshop* (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, 57-64.

H. Ishibuchi, T. Nakashima, and T. Murata. 1999. *Trans. on Systems, Man, and Cybernetics*, vol. 29, pp. 601-618.

M. Plantie, M. Roche and G. Dray. 2008. *Proceedings of the 4th DEFT Workshop* (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, 65-74.

M.A. Potter and K.A. De Jong. 2000. Cooperative coevolution: an architecture for evolving coadapted subcomponents. *Trans. Evolutionary Computation*, 8 (Jan. 2000), 1-29.

A.-P. Trinh , D. Buffoni and P. Gallinari. 2008. *Proceedings of the 4th DEFT Workshop* (Avignon, France, June 8-13, 2008). DEFT '08. TALN, Avignon, France, 75-86.

Biographical Sketch

Tatiana Gasanova studied Applied Mathematics and Computer Science in Siberian Federal University (Krasnoyarsk, Russia) with the focus on the application of evolutionary algorithms and artificial neural networks to supervised and unsupervised classification. In 2010, she received her Master Degree with the project “Elastic maps synthesis with evolutionary algorithms”. In 2011 she joined the Dialogue Systems Group as a PhD student and research assistant under the supervision of Prof. Dr. Dr.-Ing. Wolfgang Minker. Her topic continues the previous work with the main focus on text categorization and Natural Language Processing.

Nadine Glas

Télécom-ParisTech
Département TSI
Bureau DB302
37/39 rue Dareau
75014 Paris
France

nadine.glas@telecom-paristech.fr

1 Research Interests

My current research aims at developing a model of **engagement** in multi-modal interactions with **virtual agents**. To simulate engagement by the agent and try to establish engagement from the user's side, the agent needs to interact as human-like as possible. This requires representations of elements from a human mind, such as emotions, that lead to believable verbal and non-verbal behaviours. To achieve this, the agent needs to be able to manage dynamically verbal as well as non-verbal behaviour and its verbal behaviour should be generated in a similar way as other forms of behaviour.

This has led me to the following observation: In the field of NLP, dialogue systems are usually developed as independent systems to manage speech. In the field of AI, (**task-based**) dialogue systems are often merely employed as a tool without giving it further attention. This in contrast to systems that direct for example emotions and non-verbal behaviour. I believe though, that in order to allow for **multi-modal interaction** with artificial agents, dialogue systems and **cognitive systems** are intertwined concepts where the one cannot be considered without the other. I hope to incorporate such an intertwined view in a new or already existing system.

In the following sections I shall shortly discuss some points to take into account to realize this view. For the record, I do not claim that my considerations are new; there might already be systems that take into account some of these points.

1.1 Flexibility

My first point comes from the need of the agent to be able to react dynamically to the ongoing interaction as this is a crucial point in maintaining or establishing engagement by the agent and the user; Not reacting appropriately to the user's verbal and non-verbal contributions can be seen as a sign of disengagement by the agent. Moreover, dynamic reactions are necessary to directly address signs of for example disengagement or misunderstanding by the user which can arise at any moment in the interaction. Consequently, a dialogue in which the agent's contributions are completely predefined does not allow for sufficient flexibility. Even in task-oriented interactions, in which a rigid dialogue structure could be sufficient to achieve the goal of the interaction, there should be a possibility to deviate more or less from the original plan. This would allow the agent to react appropriately to unexpected changes in the ongoing interaction, offering more possibilities to engage the user, which may not only lead to a better user experience (O'Brien and Toms, 2008), but may also contribute to a successful execution of the task.

Another element that determines the required level of flexibility in a dialogue system is the domain of interaction. Task-oriented dialogue systems for example, such as Disco for Games (Rich and Sidner, 2012), are often employed for the task of talking about a certain topic. In this case, the interaction is however restricted to this very topic and the course of the interaction is completely controlled by the agent. A **chat-bot** on the other side, allows for interactions with a less restricted topic but the agent looses control over the course of the interaction. In some contexts however, we would like to interact about a particular topic but with enough freedom to react to off-topic contributions by the user. In other cases, we would like to participate in small-talk with a user but to keep the conversation going, have a topic to fall back to. This illustrates the need for a dialogue system which combines a task-based orientation with chat-based components.

A last point regarding the flexibility of a dialogue manager refers to the ability to express the same semantics in different syntactic forms. Especially in dialogue systems employed by agents this is an essential point; Not only does it ensure variability between conversations, but more important, it allows for a way to reflect personal aspects of the agent such as it's mood and personality.

1.2 Multi-Modality

A dialogue system which is capable of handling multimodal interaction differs from conventional dialogue systems at several stages. Not only text but also non-verbal behaviour consisting of gestures, movements and gaze needs to be recognized, understood, processed, synthesized and emitted. Challenges lie especially at the semantic levels of understanding and synthesizing where the

meaning of non-verbal behaviour should be interpreted in its context of dialogue history and possible accompanying speech. Bodily actions may be employed as a part of uttering something to another in an explicit manner (Kendon, 2004). But also in an implicit manner it may provide information about once intentions, interests, feelings and ideas (Kendon, 2004). Even if at this moment the automatic interpretation of non-verbal behaviour may be a long way from performing accurately, the multi-modal dialogue manager should provide room to take into account all the interpreted non-verbal parameters. Only in this way contributions by the user can be considered in its totality which may lead to a more balanced reaction by the agent.

On the production side a dialogue manager should also provide room to define verbal as well as non-verbal behaviour. The dialogue manager can generate both types of behaviour equally by, instead of formulating sentences directly, generating some form of higher-level behaviour, such as an agent's "communicative intentions". Based on interpreted information of what is going on in the interaction as well as an agent's characteristics such as its emotions and thoughts, the dialogue manager can firstly define the communicative intention that an agent's wants to express. Then only in the generation phase these communicative intentions can be expressed by mapping them to possible verbal as well as non-verbal behaviour.

2 Future of Spoken Dialog Research

- I expect that in the future dialogue systems will be more and more oriented towards multi-modal interactions. While the conviction that face-to-face dialogues are more complex than just a sequence of verbal utterances may already be spreaded widely, recent advances on the areas of computer vision, body movement recognition, eye-tracking and facial expression recognition give us more and more opportunities for such implementations.

- Further, I expect an evolving convergence between the behaviour of conversational agents and their users. While agents are not yet completely integrated in daily life and have not yet achieved the level of full realistic human-like behaviour, users will act differently towards them than towards real human-beings. However, as agents pass through different stages of integration and human-resemblence, users may find themselves adapting their behaviour constantly to the new interaction partner. I think that this period of adapting to one another may lead to previously unknown forms of dialogue in a similar way as email can be considerd as a previously unknown form of letter.

3 Suggestions for Discussion

- Usually in dialogue systems we only interpret speech. However, in face-to-face dialogues there are also non-verbal signs to interpret, and even in mediated forms of dialogue there are multi-modal parameters such as a person's prosody. Without processing these cues we might loose valuable information. Does this mean that every dialogue manager should become a multi-modal dialogue manager? Or are there still reasons to treat speech separately from other forms of behaviour?

- Above I described the need for a system that can switch between contributing relevantly to an interaction about a particular topic and elaborating (inevitably more superficially) on other topics. We might achieve this with a combination of task-based and chat-based systems. What challenges do we face in creating such a system? Do we have any other options to achieve a similar result?

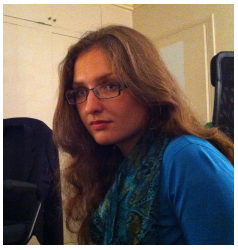
References

Kendon Adam. 2004. *Gesture. Visible action as utterance*. Cambridge University Press.

O'Brien Heather L. and Elaine G. Toms. 2008. What is user engagement? A conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6), 938–955.

Charles Rich and Candace L. Sidner. 2012. Using collaborative discourse theory to partially automate dialogue tree authoring. *Intelligent Virtual Agents*, 327–340. Springer Berlin Heidelberg.

Biographical Sketch



Since January 2013 Nadine Glas is conducting PhD research under the supervision of Catherine Pelachaud at the 'Greta-team' of Télécom ParisTech. Her broad interests have lead her through various domains. She holds a BA in Communication and Information Studies from the University of Groningen (The Netherlands) and received her Erasmus Mundus masters in Language and Communication Technologies from the university of Groningen and the University of Nancy 2 (France). She has worked as an engineer at the French labarotary LIMSI-CNRS and at a company for semantic solutions called Syllabs.

Matthew Henderson

University of Cambridge,
Department of Engineering,
CB2 1PZ, U.K.

mh521@eng.cam.ac.uk
matthen.com

1 Research Interests

Broadly, my research is in **statistical methods** for **language understanding** and **state tracking** for spoken dialog systems.

1.1 Spoken Language Understanding

In statistical spoken dialog systems, it is important that each component in the pipeline can maintain uncertainty in both its input and outputs. For example, a spoken language understanding component should use as much of the posterior distribution $\mathbb{P}(\text{Sentence}|\text{Acoustics})$ as possible, and should output an accurate approximation to the distribution over semantic hypotheses.

My previous work has looked at using discriminative methods to train robust spoken language understanding components, which decode straight from the word confusion networks computed by a speech recognition component (Henderson et al., 2012). Confusion network decoding was found to outperform the more conventional approach of using speech recognition n-best lists, which represent a more coarse approximation of the recogniser's posterior distribution over what has been said.

1.2 Dialog State Tracking

Here 'dialog state' loosely denotes all the important information to which a planner must have access when deciding what action the dialog system ought to take next. In the context of 'Let's Go', a simple bus information system, this is the value of various slots which define the information the caller is interested in (bus route, departure location, arrival location, time etc.) For the Cambridge Restaurant Information system, this includes the caller's current preferences (food=chinese, area=west etc.) and the information the caller requires (phone number, address, etc.)

Classifying the dialog state, or 'belief tracking', was the task of The Dialog State Tracking Challenge (Williams, 2012). The challenge has exposed the advantages of using discriminative methods over the previously more prominent generative techniques, including the Deep Neural Network framework that I worked on (Henderson et al., 2013).

Figure 1: Demo visualising the Cambridge Restaurant Information System, which I should be able to share at the workshop. Fully statistical implementation of a dialog system, using confusion network decoding, expectation propagation over dynamic Bayesian networks, and Gaussian processes for reinforcement learning.

Future directions in my research in dialog state tracking will include:

- Studying Dynamic Bayesian Networks whose structure (i.e. conditional independence assumptions) are learnt automatically from data
- Neural networks and dynamic Bayesian networks which take as input the results of the speech recogniser (e.g. the confusion network), and directly update the dialog state without the need for a spoken language understanding module
- Learning for flexible domains, such as ontologies created dynamically according to the current location of the caller
- Studying the impact of improved dialog state tracking on factors such as success rate in a complete dialog system

2 Future of Spoken Dialog Research

Dialog promises a natural and effective method for users to interact with and obtain information from computer systems. Traditionally, dialog systems have been deployed in call centres, but currently there is growing interest in mobile applications. Successful implementation of

a dialog system on a mobile device would allow the user to express, discuss and refine their needs through conversation with the system, as opposed to the one-off query and response model of e.g. Apple's Siri.

Multimodal applications will be a focus of research as smart phones and tablets become ubiquitous. In the context of statistical systems, this will involve increasing the action spaces to beyond speech actions to include e.g. displaying different views on screen on the system side, and gestures on the user side, while dealing with these in a probabilistic and reasoned manner. Research on incremental dialog in this context will also be interesting, allowing for well flowing conversations with the system.

Statistical methods should be at the centre of developments in dialog not only in order to provide robust systems which can cope with the uncertainty in language and speech recognition, but also to allow for the efficient and automatic learning of dialog policies without the tedious process of tweaking hand-crafted rules. There is potential for modern statistical machine learning to allow intelligent systems which can learn quickly from minimal annotation alone.

There are still multiple challenges in applying machine learning to dialog. One large one is adapting learning to domains which might change. Most schemes for learning statistical dialog systems currently require fixing the domain, and then learning domain-specific components. The future will see work towards the goal of systems which can deal with web-like scale and openness in what they can talk about.

3 Suggestions for Discussion

- Deployment of dialog systems on the web and mobile devices
- Learning for flexible and growing domains
- Automatic optimisation of incremental systems using micro-turn managers
- Reassessing the need for an intermediate stage for Spoken Language Understanding

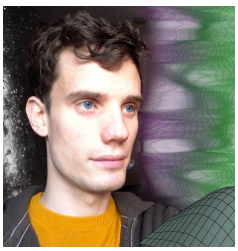
References

Matthew Henderson, Milica Gašić, Blaise Thomson, Pirros Tsiakoulis, Kai Yu and Steve Young 2012 *Discriminative Spoken Language Understanding Using Word Confusion Networks* IEEE SLT 2012

Jason Williams 2012 *A belief tracking challenge task for spoken dialog systems* NAACL HLT 2012 `http://research.microsoft.com/en-us/events/dstc/`

Matthew Henderson, Blaise Thomson and Steve Young 2013 *Deep Neural Network Approach for the Dialog State Tracking Challenge* SIGdial 2013

Biographical Sketch



After completing his undergraduate degree in Mathematics at Cambridge, Matthew did his masters in Speech & Language Processing at Edinburgh. He is now a second year PhD Student working in Steve Young's Dialog Systems group at the Engineering Department of Cambridge University. In 2013 he was awarded a Google Doctoral fellowship in Speech Technology. Outside of research, he enjoys climbing, brewing and making animations.

Casey Kennington

Universität Bielefeld
Center of Excellence Cognitive
Interaction Technology
Department of Linguistics and
Literary Studies C5-208
Universitätsstraße 25, 33615
Bielefeld, Germany

ckennington@cit-ec.uni-bielefeld.de
www.caseyreddkennington.com

1 Research Interests

There are many important components to a successful dialogue system such as ASR, linguistic processing, dialogue management, gesture recognition, dialogue act recognition, speech signal processing, monitoring belief states, among many others. These need to be combined and used in such a way that offers a positive user experience so communication can occur effectively. To that end, many of those components contribute to improving **natural language understanding** (NLU). However, there are some fundamental components that form the basis of NLU, such as the words in the utterance itself and some kind of common knowledge of how those words contribute to meaning. My PhD research has focused on NLU and what information sources contribute to understanding. Even though there are other interactive sources and behaviors that contribute to NLU and overall better communication, such as knowing when to speak, recognizing sarcasm, etc., there are *fundamental* information sources such as language, spatial context, temporal context, the other person's communicative goals, among others, that are essential to NLU.

1.1 Situational Dialogue

In (Kennington and Schlangen, 2013), we showed in a small domain that jointly using properties of objects in a shared visual context (color, shape, and spatial relations), the words and linguistic structure of an utterance, as well as knowledge of the previous utterance, were necessary to *understand* what action the user wanted the dialogue system to make, what object to take that action on, and what the resulting state of the shared visual world should be. It is clear that words and linguistics, as well as a recognition of objects in the shared space, and previous reference contributed to NLU. My more recent work, (Kennington et al., 2013; Kousidis et al., 2013), used simpler models of grounding language to a visual world, as well as incorporating eye gaze and deictic gestures.

1.2 Incrementality

Dialogue by definition is **incremental** in that an utterance itself is an incremental unit of a dialogue (Schlangen and Skantze, 2011). Further, language unfolds over time (Frazier, 1987) in incremental units which are on a finer-grained scale than sentences, or even words. Humans can perceive and understand an utterance on-line as it is being spoken. This is already a motivation for building dialogue systems that can process input incrementally, but it is also a matter of practicality; a dialogue system that continually processes new input will provide a more natural user experience. My work until this has emphasized *incremental* natural language understanding. I currently use, and will continue to use, the Inpro Toolkit (Baumann and Schlangen, 2012), which is an implementation of an incremental dialogue framework.

1.3 Future Work

Future work will involve implementing and testing our model in a more interactive setting, which will require some work in **dialogue management**. We also will move to larger domains and incorporate other linguistic information.

2 Future of Spoken Dialog Research

- Dialogue systems that can interact more naturally (not relying on fragmented turn-taking), thus causing less frustration to the human user. This requires better overall understanding from all aspects of dialogue. The less frustrated human users are with a dialogue-system, the more likely they are to use dialogue systems, which means more usable data for research in all areas of dialogue. Some of our research focuses on batch processing, not so much direct interaction, which is important, but there needs to be a balance. As we improve our dialogue systems, we always need to take time and see how things fit into a real interaction scenario.
- Incorporating various information sources should not impede real-time interactive behaviors. For ex-

ample, information about the current situation (objects in the room, gestures, eye gaze), a common knowledge base (both presumably know who a famous person is), but without fluid interaction, a human may not have patience to interact with a dialogue system. If a human doesn't treat a dialogue system like another human to some degree, some of the information sources might be lost or cause noise (i.e. the system might detect sarcasm when the human is purposefully trying to not show any emotion at all).

- People are expecting more human-like interaction with their every-day devices (e.g., ASR instead of a keyboard on a phone). We need to look at how that interaction is taking place day-to-day and how we can improve it.

- As people become more familiar and accepting to dialogue systems, they will expect them to process more complicated utterances, which will require deeper linguistic understanding.

3 Suggestions for Discussion

- *What is interaction?* A dialogue system can appear to *act* human-like but if understanding doesn't take place, is interaction really taking place?

- NLU *vs.* DM*?* Is NLU really just intra-sentential dialogue management? Are the two problems solved with similar approaches?

- *What can dialogue systems currently do well?* There are some things that computers can do better than humans, and visa-versa. What should our dialogue systems do *well*? Is there anything that a user can have high expectations of in our SDS?

- *What does a* DM *want?* Is there something other than a frame that a dialoge manager can use ("beyond the frame")?

References

Timo Baumann and David Schlangen. 2012. The InproTK 2012 Release. In *NAACL*.

Lyn Frazier. 1987. Sentence Processing: A Tutorial Review. In M Coltheart, editor, *Attention and Performance XII The Psychology of Reading*, volume XII, pages 559–586. Erlbaum.

Casey Kennington and David Schlangen. 2013. Situated incremental natural language understanding using Markov Logic Networks. *Computer Speech & Language*.

Casey Kennington, Spyros Kousidis, and David Schlangen. 2013. Interpreting Situated Dialogue Utterances: an Update Model that Uses Speech, Gaze, and Gesture Information. In *SIGdial 2013*.

Spyros Kousidis, Casey Kennington, and David Schlangen. 2013. Investigating speaker gaze and pointing behaviour in human-computer interaction with the mint.tools collection. In *SIGdial 2013*.

David Schlangen and Gabriel Skantze. 2011. A General, Abstract Model of Incremental Dialogue Processing. *Dialoge & Discourse*, 2(1):83–111.

Biographical Sketch

Casey Kennington is a PhD candidate at the University of Bielefeld, Germany, advised by Professor David Schlangen. He graduated Brigham Young University, U.S.A., in computer science, then went onto masters work in the Erasmus Mundus LCT program where he spent his first year at Saarland University, Germany, and his second year Nancy 2 University, France. He enjoys reading, running, and studying languages (now Japanese, German, and French). He and his wife, Katie, have three daughters.

Hatim KHOUZAIMI

Orange Labs / Laboratoire d'informatique d'Avignon (EA 4128)
38-40 rue du General Leclerc
92794 Issy-les-Moulineaux Cedex 9, France
339 chemin des Meinajaries 84911 Avignon cedex 9, France

hatim.khouzaimi@orange.com

1 Research Interests

Most spoken dialogue systems (SDS) available today interact with the user in a turn-taking manner. When the latter talks, the system records his voice and waits until it detects a silence before processing the request. Equivalently, when the system is speaking, the user is supposed to listen with no possibility for him to interrupt the machine. Yet, that is not the way humans communicate and early psycho-linguistic studies have shown that (Tanenhaus et al., 1995).

When we listen to somebody talking, we tend to make a partial understanding of what has been said so far before the end of the sentence and more surprisingly, we have the ability to infer the lacking piece of the information in order to understand the global meaning (DeVault et al., 2011). That is known in the literature as **incremental dialogue** (Schlangen and Skantze, 2011) and that's where my research interest lies.

Moreover, I am particularity interested in the use of **machine learning** algorithms for making SDS more efficient (Lemon and Pietquin, 2007), (Laroche, 2010). These algorithms are fed with real data extracted from real interaction with users. More specifically, my research covers the area of **reinforcement learning**.

1.1 Incremental dialogue

Disserto is a solution developed by Orange and it is used to design dialogue systems both in text and vocal mode. Disserto-based systems take the form of web-services that can be deployed in any **servlet container** (Tomcat, Jonas...). My current work is to find an optimal way for making it possible to interact with that system incrementally instead of its current traditional turn-taking manner of handling dialogue. Such a study can be useful to make other big scale systems in the industry incremental, thus making a new generation of SDS.

1.2 Reinforcement learning

So many choices and decisions have to be made while developing a dialogue system. Some of them can easily be made by the designer whereas others are not obvious at all. The best solution to address that problem is to make the system learn from a dialogue corpus formerly collected from real interactions with people; that's **offline learning**. Another variant of this approach is called **online learning** and consists on making the system learn even outside the bench test, when it is in production. It has the advantage of learning the new tendencies but it is more difficult and risky.

When the corpus is collected, each dialogue can be evaluated in different ways: by human beings and thus receiving subjective grades or automatically by extracting some features like the length of the dialogue and whether or not the user reached his goal. Based on these results, reinforcement learning provides a way for the system to learn optimal actions to take in different situations (states of the dialogue) (Sutton and Barto, 1998).

2 Future of Spoken Dialog Research

What SDS are able to do today is already quite impressive, nevertheless, they are still far from offering an ideal human-like experience and unfortunately, misunderstandings and desynchronizations are still very frequent. I think the current research efforts will lead to systems that mimic human dialogues in a way that makes them more robust and less error-prone.

Current advances in machine learning applied to dialogue are very promising (reinforcement learning and deep learning). Nonetheless, they require a big amount of data which is not easy to obtain in the field of dialogue. In some cases we can use simulation to generate data but it is not easy to determine if our model is accurate or not and this trick does not produce new useful information to learn from as real dialogs do. In my view, finding ways to address that problem is crucial and in a few years, it will be necessary to be able to get this kind of data easily on the web (for a reasonable cost or even for free).

Moreover, we have so many different spoken dialogue systems in the industry and research labs but after making mistakes and gathering experience and after the optimal ways of doing things will be identified, it will be important to have a common framework with common norms

and conventions and the software that goes with it. It will allow developers and researchers to save time (not having to do things that already have been done) but also, to create a shared experience and avoid mistakes that have already been made.

3 Suggestions for Discussion

- Plugging incremental dialogue in recent dialogue systems that have a more turn-taking oriented architecture makes it possible to save time and benefit from all the features that have been developed by designers so far. However, starting an incremental system from scratch makes it easier to make the best choices from the beginning without constraints. It would be interesting to discuss the advantages and the drawbacks of both approaches.

- One can also leverage incremental dialogue to build more robust and user-friendly multi-modal systems. For example, an avatar can show expressions while the user is speaking (e.g. nodding his head, frowning...). In my view, it's interesting to discuss the applications one can imagine that take advantage of multi-modality and incremental dialogue at the same time.

- We commonly hear that Siri had shed light on the work of researchers and developers in the field of spoken dialogue systems, making it more visible to the rest of the world. However, after everybody has played and made jokes with the assistant, only a few continue to really use it. What are the features one can expect from the next buzz-maker speaking agent that will engender a whole new generation of personal assistants?

References

David DeVault, Kenji Sagae, and David Traum. 2011. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue and Discourse*, 2:143–170.

Romain Laroche. 2010. *Raisonnement sur les incertitudes et apprentissage pour les systemes de dialogue conventionnels*. Ph.D. thesis, Paris VI University.

Oliver Lemon and Olivier Pietquin. 2007. Machine learning for spoken dialogue systems. In *Proceedings of the European Conference on Speech Communication and Technologies (Interspeech'07)*, pages 2685–2688, August.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue and Discourse*, 2:83–111.

Richard S. Sutton and Andrew G. Barto. 1998. *Reinforcement Learning, An Introduction*. The MIT Press, Cambridge, Massachusetts, London, England.

Michael K. Tanenhaus, Michael J. Spivey-Knowlton, Kathleen M. Eberhard, and Julie C. Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.

Biographical Sketch

After obtaining his engineering degree from the french school Telecom ParisTech in the field of Applied Mathematics and Computer Science and his Master's degree from the University of Paris VI in the area of Probability and Finance, Hatim worked for three years at the high frequency trading desk at the French bank Societe Generale. He then decided to begin a PhD program in the field of Machine Learning and Speech Processing which he did at Orange Labs and the University of Avignon. He is currently a first year PhD student.

Alejandra Lorenzo

Université de Lorraine
LORIA, UMR 7503
Vandoeuvre-lès-Nancy
F-54500, France

alejandra.lorenzo@loria.fr
http://www.loria.fr/~lorenzoa/index.html

1 Research Interests

Formal systems, the area of Artificial Intelligence or, in more general terms, how to bridge the gap between human and computer capabilities has always been very attractive to me. Currently, my research centers in the area of natural language processing, in particular for **dialogue systems** involving conversations with human speakers in the context of **educative games**.

1.1 About Dialog Systems and SLA

An important tenet of contemporary **Second Language Acquisition** (SLA), is that language is best learned through practice. However, language learners usually have few opportunities to use the language they are learning (because of different reasons, like shyness, lack of time of the teacher, etc.).

Because they do not intimidate the learner, computers are potentially ideal partners for language learning practice. In particular, we could use "**chatterbots**" (or "chatbots") as a tool for language practice, since they can be designed to direct conversations toward the use of a given verb tense, or a particular topic of interest for the learner.

My current work consists in developing such a chatbot in the context of the I-FLEG serious game (Amoia et al., 2012) for language learning developed for the EU funded ALLEGRO project (`http://talc.loria.fr/-ALLEGRO-Nancy-.html`). The ALLEGRO project aims to develop web services and in particular, serious games for language learning.

The main focus of my work is currently on the Understanding module (NLU which stands for Natural Language Understanding). There are various issues involved in the development of a dialog system for language learners. Particularly for the NLU module, one of the major is the issue of ill formed input: learners of a language can be expected to make more errors than native speakers. Another crucial issue is **portability**: porting an existing system to a new language or a new scenario typically involves major modifications. And finally, as we want to develop a language learning system, we need to implement some kind of error detection, since we need to provide some feedback to the learners to make them aware of the errors they make.

1.2 First Experiments

To start exploring the issues involved in the development of the NLU module, I helped in the development of the dialog system of the Emospeech project, mainly in the Understanding module. The Emospeech project aims to augment serious games with natural language (spoken and written dialog) and emotional abilities (gesture, intonation, facial expressions). The dialog system developed for this project focus on French speaking, situated conversational agents who interact with virtual characters in the context of a serious game designed to promote careers in the plastic industry. The semantic representation chosen for this system is a shallow one, based on Question-Answer characters, with limited dialogue model of the character and which focus more on retrieval of appropriate answers given a question (Rojas et al., 2012a). The dialog system and, in particular the NLU is formulated as a classification task, with a classifier for interpreting the player's phrases (Rojas et al., 2012b).

Following this same methodology, we developed a chatterbot for French language learning (Gardent et al., 2013), which we integrated in the I-FLEG serious game. In this case 2 dialoue systems where developed: one, more suitable for advanced learners, which allows free answers to the learner; and a second one, better adapted to begginers, which propose answers to the learner in the form of grammar exercises.

However, the supervised approach used in the Emospeech project requires to collect and annotate new data each time we change the language or domain. I am currently investigating possible solutions to this problem. If we consider the task of semantic role labeling as (part of) the interpretation module, the goal is to develop an automatic semantic role labeler, which would serve as interpreter. So, in such a situation, I started experimenting with unsupervised techniques to Semantic Role Labeling. In (Lorenzo et al., 2012) we propose an unsupervised approach to semantic role induction that uses a generative Bayesian model. My main motivation is to be able to use this module in different dialog systems and domains without the need to collect and annotate new data each time (as opposite to supervised methods).

In this context, we have recently proposed in (Lorenzo

et al., 2013) an unsupervised model that allows to infer latent semantic structures on top of manual speech transcriptions in a spoken dialogue reservation task. The idea behind this model is very similar to Semantic Role Labeling, except that it does not rely on any syntactic information and it exploits concepts derived from a domain-specific ontology.

Another topic that is very interesting to me, is the combination of symbolic and statistical methods. We have recently proposed in (Cerisara et al., 2013) a weakly supervised dependency parser that can model speech syntax without relying on any annotated training corpus. We replaced labeled data by a few hand-crafted rules that encode basic syntactic knowledge, and which are combined with an unsupervised Bayesian model.

Finally, I worked on the error detection issue. In this context, we focus mainly on one type of errors, namely pronouns. For that, I developed an online tool for data collection, where learners of French can do a variety of pronoun exercises. The tool is available on line (`http://talc.loria.fr/D-FLEG.html`) and it allows kind of users permits: learners and teacher. The exercises where designed by a French teacher, who, in the teacher profile, can create new exercises.

2 Future of Spoken Dialog Research

Taking into account the increasing use of Internet and speech technology, I expect that the future in this area would somehow involve deepening the knowledge about and improving the quality of the human-machine interaction.

About dialog systems in general, I always thought that in order to create a machine that behave like humans, we should first take a look at "what" we want to mimic and learn and behave as much as possible they way they learn. When we engage in a conversation, we usually use some background knowledge, we may apply some reasoning in order to infer new facts, we may also use some statistical inference. And we do it as a joint task. In the same way, I feel that not just one, but a combination of methods are needed to overcome the problems that each one present nowadays. In that sence, I like very much the work of Stuart Russell on "Unifying logic and probability".

I also think that if we want a system to learn to communicate as we do, the system is not just ment to apply some algorithm and stop learning. In this other sence, I think that the work of Tom Mitchell on "Never-Ending Language Learning" (NELL) is very promising.

3 Suggestions for Discussion

Possible topics for discussion:

- Existing unsupervised approaches used in dialog systems (for dialog act classification or semantic role labeling).
- Combination of symbolic and statistical methods, application to dialog systems.
- Portability across domains as a desired quality of a dialog system.
- Life-long spoken dialogue systems.

References

M. Amoia and T. Brétaudière and A. Denis and C. Gardent and L. Perez-Beltrachini. 2012. *A Serious Game for Second Language Acquisition in a Virtual Environment*. Journal on Systemics, Cybernetics and Informatics (JSCI) 2012.

Lina Maria Rojas-Barahona and Alejandra Lorenzo and Claire Gardent. 2012. *Building and Exploiting a Corpus of Dialog Interactions between French Speaking Virtual and Human Agents*. *LREC 2012*.

Lina Maria Rojas-Barahona and Alejandra Lorenzo and Claire Gardent. 2012. *An End-to-End Evaluation of Two Situated Dialog Systems*. *SIGdial 2012*.

Alejandra Lorenzo and Christophe Cerisara. 2012. *Unsupervised frame based Semantic Role Induction: application to French and English*. ACL/SP-Sem-MRL 2012.

Claire Gardent and Alejandra Lorenzo and Laura Perez-Beltrachini and Lina Maria Rojas-Barahona. 2013. *Weakly and Strongly Constrained Dialogues for Language Learning*. *SIGdial 2013 (Demo) Submitted*.

Alejandra Lorenzo and Lina Maria Rojas-Barahona and Christophe Cerisara. 2013. *Unsupervised structured semantic inference for spoken dialog reservation tasks*. *SIGdial 2013*.

Christophe Cerisara and Alejandra Lorenzo and Pavel Kral. 2013. *Weakly supervised parsing with rules*. *INTERSPEECH, 2013*.

Biographical Sketch

Alejandra Lorenzo is currently a third year PhD Student at LORIA Nancy grand Est, in France. She is a member of the Synalp Team where she works under the supervision of Claire Gardent and Christophe Cerisara. She was born in Argentina, where she obtained her first Masters degree in Computer Science. In 2009, she obtained a second Masters degree, after finishing the Erasmus Mundus European Masters Program in Language and Communication Technologies (LCT) at the University of Nancy 2 (France) and the Free University of Bolzano (Italy).

Yoichi Matsuyama

Department of Computer Science,
Waseda University
27 Waseda-cho Shinjuku-ku,
Tokyo, Japan

matsuyama@pcl.cs.waseda.ac.jp
www.matsuyama.tv

1 Research Interests

My research interest is models of multiparty conversation facilitation. The current research project is development of the **SCHEMA** ($\int e{:}ma$), a framework for conversational robots facilitating small talks in small groups. The main functions of the framework are: (1) Small group facilitation strategies including floors and topic maintenance (Matsuyama et al., 2013), and (2) Informative question answering for small talks.

1.1 POMDP-based Facilitation Strategies in Multiparty Conversations

We propose a facilitation robot harmonizing four-participant conversations. Four-participant conversation is the minimum unit that needs facilitation skills. In general, three is the minimum number of participants of a multiparty conversation. In such three-participant situations, back-and-forth interactions between two participants out of three primarily occur and another participant tends to be left behind, who cannot properly get floors to speak. Here, they need one more participant who helps the participant left behind to harmonize him/her with the others. Conversational robots have potentials to participate in such conversations as the fourth participant as is shown in Figure 1. In this figure, the participant C is unharmonized, and the robot is trying to approach to him. When the robot steps in the situation to help, there should be proper facilitating procedures to obtain initiatives to control conversational contexts.

There were some researches of specially situated facilitation agents in multiparty conversation. We have developed a multiparty quiz game typed facilitation system as an elderly care (Matsuyama et al., 2008) and reported the effectiveness of the existence of a robot (Matsuyama et al., 2010). Dosaka et al. developed multiparty conversation activation system in quiz task (Dohsaka et al., 2009) . Bohus et al. considered modalities including gaze, gesture and speech for facilitating multiparty conversation (Bohus and Horvitz, 2010). As for facilitation skills of group process, Kumar et al. designed dialogue action selection model based on a sociologist Bales's Socio-Emotional Interaction Categories for text based character agent (Kumar et al., 2011) . However, there is a

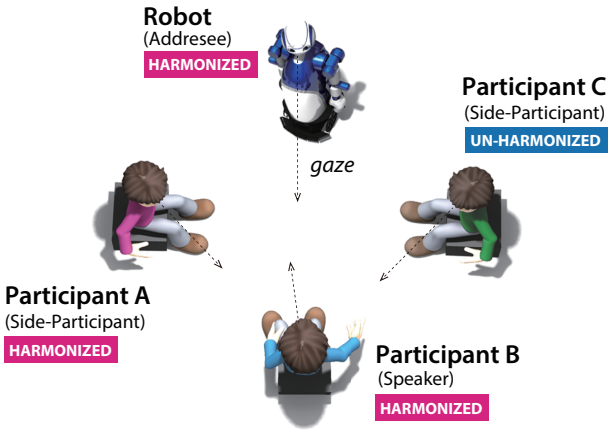


Figure 1: Four-participant conversational group. Four participants, including a robot, are talking about a certain topic. In this moment, the participant A and B are leading this conversation, who mainly keep their floors. The robot is also harmonized with A and B along a topic. C is an *unharmonized* participant, who cannot have many chances to take his floors for a while. Dashed arrows represent their gazes.

lack of deep considerations of spontaneous procedures of facilitation in multiparty settings. So we propose a conversational robot system harmonizing four-participant conversations along procedures of obtaining initiatives of topic and floor control, where all participants' chances of taking floors are supposed to be equal. These situations and procedures were modeled and optimized as the partially observable Markov decision process (POMDP) (Williams and Young, 2007). As the experimental platform, we use the SCHEMA platform we produced (Matsuyama et al., 2009).

1.2 SCHEMA QA: Informative Question Answering System

We propose the SCHEMA QA, a framework of non-task oriented question answering systems to enjoy conversations themselves. In such conversations, informative phrasing and responding skills are employed. In this

paper, we propose the enjoyable a framework that has such fundamental conversational skills including: (1) Additional responding, (2) Expressive opinion generation. In functional conversations, Grice 's Maxim of Quantity suggests that responses should contain no more information than was explicitly asked for. However, in our daily conversations, more informative productions with additional phrasing usually occur to enjoy a conversation with interlocutors, which sometimes contradicts the Grice 's Maxim. Our preliminary experiments, combination of passive response and spontaneous phrasing with own opinions indicates effectiveness to entertain conversations (Matsuyama et al., 2011). The opinions in our current system are extracted from a large number of reviews in the web, and ranked in terms of contextual relevance, length of sentences, and amount of information. Our experimental results show that both additional responding and expressive opinions mechanisms are effective to promote users' enjoyment and interests.

2 Future of Spoken Dialog Research

We need to go far beyond the recent success of industrial spoken dialogue systems, which are mostly simple factoid-typed question answering. One of possible directions I believe is to build computational models of deeper discourse understanding with multimodal inputs.

3 Suggestions for Discussion

- **Beyond Success of Industrial QA Systems**
  How can we go far beyond the current success of industrial QA systems, such as Apple's Siri and IBM's Watson?

- **Cognitive Modeling of Conversational Process**
  I believe that spoken dialogue systems and conversational robots can be promising tools to pursue human cognitive model of conversation. Researcher of dialogue systems can take more advantages of recent neural cognitive science and social cognitive science, and contribute them from computational sides.

References

Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, page 5. ACM.

Kohji Dohsaka, Ryota Asai, Ryuichiro Higashinaka, Yasuhiro Minami, and Eisaku Maeda. 2009. Effects of conversational agents on human communication in thought-evoking multi-party dialogues. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–224. Association for Computational Linguistics.

Rohit Kumar, Jack L Beuth, and Carolyn P Rosé. 2011. Conversational strategies that support idea generation productivity. In *in Groups, 9 th Intl. Conf. on Computer Supported Collaborative Learning, Hong Kong 160 and Rosé, 2010a) Rohit Kumar, Carolyn P. Rosé, 2010, Conversational Tutors with Rich Interactive Behaviors that support Collaborative Learning, Workshop on Opportunity*. Citeseer.

Yoichi Matsuyama, Hikaru Taniyama, Shinya Fujie, and Tetsunori Kobayashi. 2008. Designing communication activation system in group communication. In *Humanoid Robots, 2008. Humanoids 2008. 8th IEEE-RAS International Conference on*, pages 629–634. IEEE.

Yoichi Matsuyama, Kosuke Hosoya, Hikaru Taniyama, Hiroki Tsuboi, Shinya Fujie, and Tetsunori Kobayashi. 2009. Schema: multi-party interaction-oriented humanoid robot. In *ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies: Adaptation*, pages 82–82. ACM.

Yoichi Matsuyama, Shinya Fujie, Hikaru Taniyama, and Tetsunori Kobayashi. 2010. Psychological evaluation of a group communication activation robot in a party game. In *Eleventh Annual Conference of the International Speech Communication Association*.

Yoichi Matsuyama, Yushi Xu, Akihiro Saito, Shinya Fujie, and Tetsunori Kobayashi. 2011. Multiparty conversation facilitation strategy using combination of question answering and spontaneous utterances. In *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pages 103–111. Springer.

Yoichi Matsuyama, Iwao Akiba, Akihiro Saito, and Tetsunori Kobayashi. 2013. A framework for conversational robots facilitating four-participant groups. In *Proceedings of the SIGDIAL 2013 Conference: The 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page accepted. Association for Computational Linguistics.

Jason Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Biographical Sketch

Yoichi Matsuyama is currently a researcher and a Ph.D candidate at the Perceptual Computing Group, Department of Computer Science, Waseda University in Japan. He received B.A. in cognitive psychology and media studies and M.E. in computer science from Waseda University in 2005 and 2008 respectively. He was a research associate of Department of Computer Science from 2010 to 2012. He was also a committee member of ACM SIGGRAPH Asia 2009, leading an international student volunteers team. His latest project is the SCHEMA, an embodied conversational robot framework, specifically its conversational strategies and question answering. He also has a background of design. The SCHEMA's styling and mechanical design is one of his work.

Raveesh Meena

KTH Royal Institute of Technology
Department of Speech, Music and Hearing
Lindstedtsvägen 24, 10044 Stockholm
Sweden

raveesh@csc.kth.se
http://www.speech.kth.se/~raveesh/

1 Research Interests

My research interests lie in using **machine learning** approaches for **modeling human-like conversational behavior** in dialogue systems. More specifically, I am interested in applying Conversational Analysis techniques to analyze human-machine interaction data; identify issues or patterns that are relevant for enhancing the usability of the system; build data-driven models of human dialogue strategies for resolving these issues; implement such models in a dialogue system and verify their utility through **user evaluations**.

A key aspect of my approaches is the use of ***bootstrapping*** procedures in designing dialogue systems. In this methodology more and more advanced versions of a dialogue system are developed iteratively. The data collected in the first iteration is used to train an improved model. The model is then implemented in the same system and evaluated through users. The user interaction data could then be used for further improving the model. This method provides an alternative to the Wizard-of-Oz simulations for data collection or human-human corpora for training models, both of which have certain limitations.

1.1 A data-driven model for timing feedback

Providing feedback at appropriate occasions during an interaction is vital for a system to indicate continued attention to the user without disrupting the flow of the interaction. In a recent work (Meena et al., in press) we have used the aforementioned bootstrapping procedure to build a data-driven model of what we call—Response Location Detection (RLD). In the first iteration, we built a dialogue system that can perform Map Task with users. Since it is not feasible to build a system with fully automated speech understanding, we introduced a trick. The users were asked to draw the route using the mouse cursor as they described it to the system. They were told that it was for logging purpose. However, the system used the current mouse position to make calculated guesses about what the user might be talking about. The system used random values for silence threshold to detect feedback relevant places in user's speech. The system selected (partly randomly) from a set of responses such as clarification, guess, repeat, and acknowledgement. Using this naïve model 50 human-computer Map Task interactions were collected. Next, we used this corpus to train a sophisticated model of RLD, which on having detected a silence of 200 ms in user's speech decides whether the system should *respond* or *hold*. A Naïve Bayes algorithm trained on automatically extractable syntactic, prosodic and contextual features from user's speech preceding a silence, was able to make these decisions with an accuracy of 82%. The model was implemented in the original Map Task dialogue system and evaluated through users. The subjective ratings suggest that the system with trained model produces less disruptive interactions in contrast to the system with the näive model.

1.2 A data-driven model for semantic interpretation

In an earlier work (Meena et al., 2012) we have applied machine learning for the task of Spoken Language Understanding (SLU). The proposed method is a novel application of Abney's *chunking parser* (Abney, 1991) in automatic semantic interpretation of verbally given route instructions to a robotic agent. Given a speech recognized hypothesis (e.g., *turn left after the church*) our method first *chunks* the hypothesis into basic domain concepts (e.g. ACTION, DIRECTION, ROUTER, and LANDMARK). In the next stage, the basic concepts are first assigned more specific concepts (e.g., TURN, LEFT, AFTER, and CHURCH), and then assigned attributes (e.g. *direction, landmark*) and neighboring concepts as attribute-value thereby *attaching* structurally related concepts (e.g., TURN→LEFT, AFTER→CHURCH, where the arrows suggest that context—left or right—of the argument concepts in the semantic representation). Automatic semantic interpretation was thus formulated as a classification task where domain concepts and structural relationships among these concepts are learned from corpus. For learning we used a Perceptron learner, and a human-robot route instruction corpus for training. The model's performance on speech recognized results highlighted the two stage method's robustness in han-

dling speech recognition errors during semantic interpretation of verbally given route directions.

2 Future of Spoken Dialog Research

For many in the spoken dialogue research community, the ultimate goal is to model human-human dialogue as closely as possible. Despite recent progress, we are still far from the goal. To improve current systems, we need both a better understanding of the phenomena of human interaction, better computational models and better data to build these models. I believe it is worthwhile to ask what aspects of human conversational behavior are essentially required in a dialogue system. Also, it is unclear how to map models of human-human interaction onto dialogue systems. Therefore, understanding of realistic human-machine interactions is vital for making tangible progress in the development of dialogue systems. Models trained on realistic interactions would be more resilient to scenarios arising in actual user interactions. However, obtaining realistic data in the absence of fully functional dialogue systems is like a chicken-and-egg situation. Bootstrapping procedures (besides Wizard-of-Oz methodologies) offer an excellent platform for collecting realistic data. The iterative stages in bootstrapping provide for training a model and evaluating it, thereby providing realistic insights into issues relevant to enhancing a system's usability. I believe in coming years bootstrapping procedures will gain momentum in the dialogue system development community.

3 Suggestions for discussion

- Dialogue system applications: are there any takers?
- Let us identify human conversational behaviors that are critical for dialogue systems' acceptability and usability.
- Let us discuss the state-of-the-art in incremental speech processing and speech generation.
- Spoken language understanding.

References

Meena, R., Skantze, G., & Gustafson, J. (2012). *A Data-driven Approach to Understanding Spoken Route Directions in Human-Robot Dialogue*. In *Proceedings of Interspeech*. Portland, OR, US.

Meena, R., Skantze, G., & Gustafson, J. (in press). *A Data-driven Model for Timing Feedback in a Map Task Dialogue System*. To be published in *14th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGDial*. Metz, France.

Biographical Sketch

Raveesh Meena is a graduate student at KTH, Stockholm. His research work is under the supervision of Gabriel Skantze and Joakim Gustafson. In his research, Raveesh explores application of machine-learning on modeling human-like conversational behavior in dialogue systems.

Before coming to KTH, Raveesh studied Language Science and Technology, at the Saarland University, Saarbrücken, Germany, where he obtained his M.Sc. with specialization in Computational Linguistics (July 2010). Raveesh got engaged in the field of human-machine interaction during his student research assistantships in the Language Technology lab at the German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken. He has four years of industry experience in software engineering. He has a Bachelor's degree in Information Technology from the Indian Institute of Information Technology, Allahabad, India (in 2003). In his spare time he likes to be outdoors, hiking or cycling. He loves to cook, likes traveling, and meeting new people.

Christopher M. Mitchell

Department of Computer Science
North Carolina State University
Raleigh, NC, USA

cmmitch2@ncsu.edu
http://www4.ncsu.edu/~cmmitch2

1 Research Interests

My research interests focus on machine learning approaches to tutorial dialogue systems. Specifically, I investigate techniques for learning dialogue management strategies from human-human tutoring corpora. I am also interested in the ways in which humans adapt to each other in dialogue, and the implications this adaptation might have for the development of an automated dialogue system. My work toward these goals has been conducted to date within the JavaTutor project, which aims to build a fully automated mixed-initiative task-oriented tutorial dialogue system for introductory computer science with both cognitive and affective adaptation to the user. To date, my research on this project has dealt primarily with studying both effective patterns in tutorial dialogue and lexical convergence in dialogue.

1.1 Studying Effective Tutorial Dialogue

A major goal of tutorial dialogue research is to learn effective dialogue management strategies from data. Toward that end, I have assisted in the collection of a sizable corpus (about 50,000 utterances across 380 interactions) of human-human task-oriented tutorial dialogue. In collaboration with others in my research group, I developed a dialogue act annotation scheme that was applied to portions of the corpus. A preliminary analysis investigated correlations between these dialogue acts and session-level outcomes such as learning gains, affective outcomes such as confusion and frustration, and student characteristics such as incoming knowledge level and domain-specific self-efficacy (Mitchell et al., 2012a). We found several unigrams and bigrams of dialogue acts that were significantly negatively correlated with desirable tutorial outcomes. These findings show promise for learning tutorial dialogue strategies in a data-driven way.

1.2 Convergence and User Adaptation

Convergence, the phenomenon of humans becoming more similar in their lexical, prosodic, and multimodal behaviors over time, has been widely studied, both within the domain of tutorial dialogue and in other domains. To better understand the role of convergence within tutoring, I have examined lexical convergence in the JavaTutor corpus (Mitchell et al., 2012b). The results indicate a longitudinal trend: users were more likely to reuse their partners' words as they engaged in more dialogues together, with a significant increase observed between the first and sixth tutoring session. Several measures of convergence were also predictive of specific aspects of both dialogue success and user affect. For example, students who reused tutor words at higher rates reported that the tasks were less mentally demanding, and tutors who reused student words at higher rates were found to be less effective at producing learning gains. These results highlight the potential for applying convergence analysis to create more effective tutorial dialogue system adaptation.

1.3 Turn-Taking in Tutorial Dialogue

In tutorial dialogue, a student completes a task for the purpose of learning a particular concept or skill while a tutor monitors the student's progress and intervenes in the problem-solving process as necessary to provide remediation, feedback, or to expand upon the concepts being learned. In tutorial dialogue, failing to provide helpful feedback to a student who is confused may lead to decreased learning (Shute 2008) or to disengagement (Forbes-Riley and Litman 2012), while providing tutorial feedback or interventions at inappropriate times could also have a negative impact on the outcome of the dialogue (D'Mello et al. 2010). In order to learn effective timings of tutor moves from the JavaTutor corpus, I created a representation of turn taking in this corpus as a Markov Decision Process and learned a tutor turn-taking policy (Mitchell et al., 2013a). This policy recommended continual tutor engagement throughout the problem-solving process, as well as avoiding making multiple tutor moves in a row. I evaluated the quality of various features used in the state representations for this Markov Decision Process using both an existing cumulative reward metric and a new *Separation Ratio* metric that estimates the importance of each decision point (Mitchell et al., 2013b).

2 Future of Spoken Dialog Research

I think a promising line of dialogue research lies in learning about and adapting to a user in the same ways that humans do. Two areas with significant potential are predicting and adapting to knowledge and skill levels of the user, and affect detection and adaptation. I believe adaptive linguistic choices have the potential to have a major impact on both the usability of a system and task success within that system, and further exploring the ways in which linguistic adaptation impacts human-human dialogue is a very promising direction for future research. In addition, for task-oriented domains, such as technical support or problem-based tutoring, being able to assess the skills possessed by the user will be fundamental in providing helpful and efficient assistance for users of all levels of expertise. Detecting and adapting to user affect will also be important in coming years as a supplement to deep natural language processing and plan recognition, particularly as user expectations regarding the intelligence of dialogue systems steadily rise.

3 Suggestions for Discussion

- *Challenges presented by mixed-initiative systems*: When should a system wait for input from the user and when should it intervene, especially in a task-oriented domain? How important is the timing of interventions in a system with relaxed turn-taking, in terms of impact on dialogue success?
- *The role of user expectations in the success of an interaction with an automated dialogue system*: Do dialogue systems need to change their behavior based on what they believe the user expects the system to be able to do? Given steady advances in the state of the art, will dialogue models learned from human-computer corpora today be valid in the future?
- *Efficient annotation of large corpora:* As the majority of dialogue research has become data-driven, how can we develop annotation schemes that capture the rich information present in dialogue while still allowing for large corpora to be tagged efficiently and reliably? What role might automated annotation play in the efficient annotation of these corpora?

References

Sidney D'Mello, Andrew Olney, Natalie Person (2010). Mining Collaborative Patterns in Tutorial Dialogues. *Journal of Educational Data Mining*, 2(1), 1–37.

Kate Forbes-Riley and Diane Litman (2012). Adapting to Multiple Affective States in Spoken Dialogue. In *Proceedings of the 13th Annual SIGDIAL Meeting on Discourse and Dialogue*, 217–226.

Christopher M. Mitchell, Eun Young Ha, Kristy Elizabeth Boyer, James C. Lester (2012a). Recognizing Effective and Student-Adaptive Tutor Moves in Task-Oriented Tutorial Dialogue. In *Proceedings of the Intelligent Tutoring Systems Track of the 25th International Conference of the Florida Artificial Intelligence Research Society*, 450-455.

Christopher M. Mitchell, Kristy Elizabeth Boyer, James C. Lester (2012b). From Strangers to Partners: Examining Convergence within a Longitudinal Study of Task-Oriented Dialogue. In *Proceedings of the 13th Annual SIGDIAL Meeting on Discourse and Dialogue*, 94-98.

Christopher M. Mitchell, Kristy Elizabeth Boyer, James C. Lester (2013a). A Markov Decision Process Model of Tutorial Intervention in Task-Oriented Dialogue. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*.

Christopher M. Mitchell, Kristy Elizabeth Boyer, James C. Lester (2013b). Evaluating State Representations for Reinforcement Learning of Turn-Taking Policies in Tutorial Dialogue. To appear in *Proceedings of the 14th Annual SIGDIAL Meeting on Discourse and Dialogue*.

Valerie J. Shute (2008). Focus on Formative Feedback. *Review of Educational Research*, 78(1), 153–189.

Biographical Sketch

Christopher M. Mitchell is a Ph.D. student at North Carolina State University, where he is advised by Kristy Elizabeth Boyer and James C. Lester. Chris holds an M.S. in computer science, and a B.S. in computer science with minors in cognitive science and mathematics from North Carolina State University. His research interests lie in computational linguistics, dialogue systems, and intelligent tutoring systems.

Aasish Pappu

Ph.D. Candidate
Language Technologies Institute
Carnegie Mellon University

aasish@cs.cmu.edu
www.cs.cmu.edu/~apappu

1 Research Interests

I find all aspects of dialog systems fascinating. I have worked on spoken dialog interfaces for mobile robots, speech interfaces deployable on the web, teaching robots through dialog and situated multiparty interaction agents.

1.1 Situated Multiparty Interaction

Human-human conversations adapt to the dynamics of the environment, in order to make the communication between people easier. The dynamics of a conversation includes the topic, speakers, hearers and the floor. Floor management in a two-way conversation a.k.a. dialog is distinct from multi-way conversation. People show efficient floor management skills either two-way or multi-way conversations. When a robot or an artificial agent is involved in such situations, one could not expect similar efficiency in the conversation. (Traum, 2004) have highlighted three major aspects of a multiparty conversation that needs to be addressed if an agent is involved. (1) Detection of participant roles (2) Interaction Management (3) Grounding. Based on these aspects, in our recent work (Pappu et al., 2013), we have looked at three research questions (a) How does an agent determine the roles in an interaction? (b) When is it appropriate to take/release the floor? (c) How does the agent ground its intentions and inferences about the floor?

To detect addresses in a multiparty conversation, (Akker and Traum, 2009) have done comparative analysis between rule-based methods and statistical methods such as bayesian networks. They found that rule-based methods are comparable to the statistical methods and we employ the rule-based method in our work. (Nakano and Ishii, 2010) and (Bohus and Horvitz, 2009) found that gaze patterns and gestures can greatly improve the addressee detection performance. In our work we use the agent's gaze to ground the agent's belief about who is being addressed. Gaze along with spoken feedback helps especially with a heuristic turn-taking policy.

In (Pappu et al., 2013) we introduced a new framework based on Ravenclaw/Olympus architecture for multiparty interactions. This framework addresses above mentioned research questions. (1) We use Microsoft kinect sensor for audio and skeletal input streams to help the agent detect the participant roles. (2) The agent makes floor management decisions based on a expert policy (defined by the agent developer) and the interaction state. (3) Finally, the agent responds to the users with a talking head to communicate its belief about who has the floor, or who is expected to take the floor? Few of the details of our approach is described in the following subsections, for the rest please refer to our (Pappu et al., 2013) paper.

1.2 Multi-user Awareness

A typical multiparty conversation framework has two fundamental requirements: (i) When to take or release the floor (ii) Who is the current speaker?Our framework fulfills these requirements by tracking and updating the skeletal positions of the participants using the Kinect sensor. The frameworkallows both dialog and multiparty conversation without having to customize the agent for each situation. An agent decides whether it is involved in a dialog or a multiparty interaction based on the number of skeletons that were detected. Once an interaction is established, each user's input is associated with its corresponding skeleton.

In order to verify the effectiveness of the multi-user awareness method, we have conducted user studies involving 12 multiparty conversations. Each conversation was carried out by two human subjects and the agent. The subjects were asked to schedule a common activity on a university campus with the help of the agent. During this experiment, the agent associated the speech with correct skeleton with an accuracy of 81%. This was measured by manually labeling each utterance with the corresponding subject. Our experiments suggest that kinect's skeletal and auditory information are reliable to detect participant roles in a multiparty interaction.

1.3 Multiparty Conversation Manager

Typical dialog agent has a dialog manager that knows what information is needed from the user, and what information is provided by the user. In a mulltiparty situation, the agent needs a conversation manager to keep track of information flow from multiple users. We have two major requirements for this conversation manager (i) It should support both dialog and multiparty situations (ii) Compatible with existing dialog applications based on Ravenclaw/Olympus framework.

In our framework, the conversation manager (CM) has access to all on-going dialogs. Each dialog context is in turn managed by a dialog manager. Therefore, the CM interacts with more than one dialog manager. The CM mediates between dialog managers and the input/output channels of the system. Depending on the individual dialog acts and the dialog states, CM takes an action. For example, DM-A and DM-B both would like to request and are in same dialog state, then the CM would make a joint request to the users. Otherwise, the CM decides its action based on an expert defined policy. For example, the policy could help DM decide which DM has got the priority to request information based on different parameters such as turns-taken, dialog history, recognition history etc.

1.4 Addressee Transparency

In a dialog setting, non-verbal cues could enhance the interaction but may not be essential for the interaction. However in a multiparty setting, non-verbal cues such as gaze are essential to determine the floor-holder at any given time. To address this necessity, we introduced a 2D talking head in our framework that helps convey which of the participants is being addressed by the agent.

We conducted both subjective evaluations to study the necessity of non-verbal cues in a multiparty setting. We have conducted 12 multiparty interactions, out of them 6 are with talking head and 6 without it. At the end of each interaction subjects were asked to rate their conversation on likert scale for different subjective questions. Subjects rated the interaction with talking head as 3.83/5 as opposed to 2.75/5 for one without it ($p < 0.01$). We observed promising improvement with respect to number of turns to complete a task when the talking head is in place. We found that both state and addressee grounding is essential for an efficient interaction.

2 Future of Spoken Dialog Research

- I believe that the dialogue research will help improve the lives of physically challenged with growing efficiency in the spoken dialog interactions.
- The systems should be made more robust and aimed to be tested in wild with real users like other areas of language and speech technologies have already been doing for a while.

3 Suggestions for Discussion

- Rapid prototyping of the dialog systems for quicker data collection.
- Making challenging games out of inherent speech recognition issues in a dialog system.
- How to break out of the traditional platforms of dialog research e.g., situated agents, telephones.

References

Rieks Akker and David Traum. 2009. A comparison of addressee detection methods for multiparty conversations.

D. Bohus and E. Horvitz. 2009. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the 2009 Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 244–252.

Y. Nakano and R. Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 139–148. ACM.

A. Pappu, M. Sun, S. Sridharan, and A. Rudnicky. 2013. Situated multiparty interaction between humans and agents. In *Proceedings of the 2013 Human Computer Interaction International*.

David Traum. 2004. Issues in multiparty dialogues. *Advances in agent communication*, pages 1954–1954.

Biographical Sketch

Aasish Pappu is currently a PhD Candidate and Research Assitant at the Language Technologies Institute, CMU at Pittsburgh under the supervision of Dr. Alex Rudnicky. He obtained his BTech degree in Information Technology from Indian Institute of Information Technology, Allahabad, India. Besides research, his interests include photography, drawing, languages and poetry.

[JongHo Shin]

[kt]
[17 Woomyun-dong, Seocho-ku, Seoul,
South Korea. Zip: 137-792]

[jhs.shin@kt.com]

1 Research Interests

My research interests lie in the optimization of the **spoken dialog strategies**, data-driven **user modeling and user simulation**, and **statistical inference mechanisms** under uncertainty. More specifically, analysis of interaction data and domain knowledge is used for designing the optimized dialog strategies, which eventually **enhance the user satisfaction and the system performance**. In my research work, the user models and user simulations are employed for customizing the speech-enabled multimodal interface to suit user needs. And, inference mechanisms such as **dynamic Bayesian network** and **Partially Observable Markov Decision Process** are adopted for the prediction of user states and for the guidance of system actions. Hereafter, I aim to apply my research output to diverse speech-enabled applications in the domains of IPTV, mobile devices, semantic search, and recommendation.

1.1 Speech-to-Speech Translation Project

My PhD work was part of the development of a speech-to-speech (S2S) translation system between Farsi-speakers and English-speakers in the medical domain (the "SpeechLinks" project of DARPA). And, I have worked on modeling user behaviors from spoken interactions and optimizing system actions with regard to the user models (Shin, Georgiou and Narayanan, 2006; Georgiou et al., 2006). I have attempted to address overcoming spoken errors with user behavior models, and have come up with three user types, called "accommodating," "normal," and "picky" in error conditions. By utilizing an inference mechanism, dynamic Bayesian network, the S2S translation system predicted user types in error conditions and generated appropriate real-time feedbacks to the targeted user types in those error conditions. This real-time feedbacks improved system performance by reducing error rates, and enhance user satisfaction (Shin, Georgiou and Narayanan, 2010). In addition, in-depth analysis of user behaviors and system responses has been done in a setting of multimodal interactions (Shin, Georgiou and Narayanan, 2007) and in a setting of spoken dialogs (Shin, et al., 2002). The analysis output indicates higher rates of user repeat and rephrase behaviors under error conditions, and higher concept transfer rate in the setting of multimodal interface. Further, longitudinal user behavior changes (e.g., user adaptation to system response over time) have been analyzed in a setting of multimodal interactions (Shin, Georgiou and Narayanan, 2012).

1.2 Spoken Dialog Research for IPTV, Mobile Devices, Semantic Search, and Recommendation

In my current research project, I attempt to produce optimal spoken dialog policies under uncertain error conditions, by employing Partially Observable Markov Decision Process (POMDP). POMDP has been actively used as a framework for generating optimized spoken dialog strategies in diverse conditions, which models optimal spoken dialog paths effectively (Young, Gasic, Thomson and Williams, 2013). The current project incorporated a special case of the POMDP management: the length of spoken dialog interaction turns is short, and the POMDP dialog management deals with both ASR errors and concept errors from language understanding. Further, developing user simulation and error model is the key in my research for the optimization of system actions with the POMDP dialog management. In this regard, the co-adaptation framework is considered to be used for building a self-evolving spoken dialog system (Chandramohan, Geist, Lefevre and Pietquin, 2012). The current project rather expands its application to semantic search, multimedia recommendation, and user profiling with the models. The status of the current project is in early stage; analyzed the IPTV domain, developed a prototype spoken dialog system, and collected sample data.

2 Future of Spoken Dialog Research

There are three trends of information technology for the near future -- mobility, big data, and clouds. In consideration of these trends the field of dialog research will lie in the following:

- Speech-enabled interface for the wearable devices, such as smart watch or the Google glasses
- Machine-learned optimal dialog strategies, precisely customized to tasks, domains, and users
- Location / cloud-based speech services, such as a location-aware speech-to-speech translation service

This generation of young researchers will be able to accomplish most of the above research efforts in the near future. This is due to the rapid growth of mobile software technologies, open-source community, and big data technologies, which eventually expedite the considerable advance in the dialog research. For this accomplishment there are a few questions to be addressed first: how to collect appropriate, meaningful big training / testing data for learning spoken dialog policies; what aspects of mobility are considered important and how those are extracted for the design of the optimal dialog strategies; how to design effectively an optimal speech-enabled interface service on cloud.

3 Suggestions for Discussion

- Parameter learning and adjustment for building statistical models and user simulations, especially when there are small amount of available data for training and testing.
- Speech or multimodal interface for the heterogeneous mobile platforms and architectures: adaptation of intuitive speech-enabled interface across diverse platforms and architectures. And, what is the core competency in the distributed computing environment?
- Spoken dialog systems or speech-command systems embedded in the wearable intelligent devices: customization of spoken dialog strategies to each wearable device.

References

JongHo Shin, Panayiotis G. Georgiou, and Shirikanth S. Narayanan. 2006. *User Modeling in a speech translation driven mediated interaction setting*, Proceedings of the International Workshop on Human-Centered Multimedia (HCM), *Santa Barbara, USA*.

JongHo Shin, Shrikanth S. Narayanan, Laurie Gerber, Abe Kazemzadeh and Dani Byrd. 2002. *Analysis of user behavior under error conditions in spoken dialogs*, Colorado, USA.

JongHo Shin, Panayiotis G. Georgiou and Shrikanth S. Narayanan. 2007. *Analyzing the multimodal behaviors of users of a speech-to-speech translation device by using concept matching scores*. Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP), Crete, Greece.

JongHo Shin, Panayiotis G. Georgiou and Shrikanth S. Narayanan. 2010. *Towards modeling user behavior in interactions mediated through an automated bidirectional speech translation system*, Computer Speech and Language, 24(2):232-256.

JongHo Shin, Panayiotis G. Georgiou and Shrikanth S. Narayanan. 2012. *Enabling Effective Design of Multimodal Interfaces for Speech-to-Speech Translation System: An Empirical Study of Longitudinal User Behaviors over Time and User Strategies for Coping with Errors*, Computer, Speech, and Language, 27(2):554-571.

Panayiotis G. Georgiou, Abhinav Sethy, JongHo Shin and Shrikanth S. Narayanan. 2006. *An English-Persian automatic speech translator: Recent developments in domain portability and user modeling*, Proceedings of International Conference on Intelligent Systems and Computing (ISYC), Ayia Napa, Cyprus.

Steve Young, Milica Gasic, Blaise Thomson and James Williams. 2013. *POMDP-based Statistical Spoken Dialogue Systems: a Review*, Proc IEEE, 101(5):1160-1179.

Senthilkumar Chandramohan and Matthieu Geist and Fabrice Lef\`evre and Olivier Pietquin. 2012. *Co-adaptation in Spoken Dialogue Systems*, Proceedings of the Fourth International Workshop on Spoken Dialog Systems (IWSDS 2012), Paris, France.

Biographical Sketch

JongHo Shin is a Research Scientist at Advanced Institute of Technology of kt (Korea Telecom) Seoul, South Korea. His interests include data-driven spoken and multimodal dialog managements, statistical user modeling, and user simulation. He holds a PhD and Masters in Computer Science from University of Southern California (USA). He has published around 10 peer-reviewed journal and conference papers and was awarded the best student paper at InterSpeech 2002 (second place) and IEEE MMSP 2007, respectively. Prior to Korea Telecom, JongHo participated in, as a research assistant, the USC-SpeechLinks project and the ICT-MRE project.

Dirk Schnelle-Walka

Technische Universität Darmstadt
Telecooperation
Hochschulstraße 10
64289 Darmstadt

dirk@tk.informatik.tu-darmstadt.de
www.tk.informatik.tu-darmstadt.de/
people/dr-dirk-schnelle-walka

1 Research Interests

My research interests lie in the area of *Collaborative User Interfaces for Smart Spaces*. Generally these interfaces are multimodal with a special focus on voice as one the major modalities.

For me these interfaces should have a *good user experience* on the one hand and a *good developer experience* on the other hand. Dialog systems that are easy to implement most likely rely on a finite state machine or a frame based dialog system but usually result in rigid dialogs. Dialogs that appear to be more natural are harder to implement. For instance the Information State Update Model introduced by Larsson and Traum (Larsson and Traum, 2000) has the potential for very flexible and natural dialogs, but for the developer it is almost impossible to consider all variations of the defined update rules. Here, an approach to dialog modeling is needed which helps to minimize the developer's efforts while keeping the potential to create more natural dialogs.

Therefore, we settle upon existing standards like the W3C standard for Multimodal Architecture and Interfaces (Bondell et al., 2012) (MMI) and analyzed its expressiveness in comparison to existing architecural considerations of the past decades (Schnelle-Walka et al., 2013b).

The architecture proposed by the W3C decomposes a multimodal application into a nested structure of *interaction managers* (IM) for dialog control and *modality components* (MC) for in- and output. An application is conceived as a set of control documents expressed e.g. in SCXML (Barnett et al., 2012) for the interaction managers and a set of presentation documents with modality-specific markup for the modality components. A topmost root controller document describes the global dialog and instantiates modality components as required. Each modality component can, in turn, again be an interaction manager, handling more fine granular concerns of dialog control, such as error correction or even sensor fusion/fission.

In order to support the ISU approach, the dialog modeling IM ought to support grounding and reasoning for the information state (Fodor and Huerta, 2006). SCXML only features XPath and ECMAScript as embedded scripting languages, both lacking these capabilities. To solve this issue, we introduced Prolog as a scripting language for SCXML documents (Radomski et al., 2013), similar to the approach by Kronlid and Lager (Kronlid, 2007). Transitions and conditions are guarded by Prolog queries, enabling application developers to employ grounding and reasoning. Other language features are available to establish a-priori knowledge or introduce new facts as the dialog progresses. Eventless transitions, only guarded by conditions, are used to apply the update rules from ISU.

This way we are able to offer the expressiveness of ISU with dialog manager that utilizes a state machine at its heart. Further extensions to embed Trindikit enables us to even support plan based dialog managers. We do not plan to extend this for other classes of dialog managers that we described in (Schnelle-Walka and Radomski, 2012; Schnelle-Walka and Radomski, 2013) to probabilistic dialog managers (Williams and Young, 2007) since we believe that they are hard to handle for application developers.

2 Future of Spoken Dialog Research

I believe that spoken dialog systems will be a fundamental part of our lives. However, voice will be only one among other modalities. Voice input and output are not suitable in all situations but they will be one of the most important modalities, either used alone or augmenting the use of other modalities. We will make use of *conversational agents* to help us accomplishing things. A study in (Schnelle-Walka et al., 2013a) showed that even command&control dialogs have the potential to be perceived as a dialog. Hence, a better understanding of context is needed as well as enabling the use of the modalities best suited for the current goal. From a developer's point of view the development of such applications must be as easy as possible, allowing them to reuse existing knowledge. Here the development of a common programming concept is needed and means for rapid application development and testing.

3 Suggestions for Discussion

For me the following topics would be interesting to discuss:

- Application development: What are the requirements to ease the development process of spoken dialog systems while providing a good user experience?
- Standardization: Which are the most appropriate standards that would application developers of spoken dialog systems to reuse their knowledge?
- Tooling: Almost everybody starts from scratch, reinventing the wheel and falling into the same traps. In fact, most of the projects die after the researcher left for another job. How could we build up a toolbox that keeps maintained?

References

Barnett, J., Akolkar, R., Auburn, R., Bodell, M., Burnett, D. C., Carter, J., McGlashan, S., Lager, T., Helbing, M., Hosn, R., Raman, T., Reifenrath, K., and Rosenthal, N. (2012). State chart XML (SCXML): State machine notation for control abstraction. W3C working draft, W3C. http://www.w3.org/TR/2012/WD-scxml-20120216/.

Bondell, M., Dahl, D., Kliche, I., Larson, J., Porter, B., Raggett, D., Raman, T., Rodriguez, B. H., Selvari, M., Tumuluri, R., Wahbe, A., Wiechno, P., and Yudkowsky, M. (2012). Multimodal Architecture and Interfaces. W3C recommendation, W3C. http://www.w3.org/TR/2012/REC-mmi-arch-20121025/.

Fodor, P. and Huerta, J. M. (2006). Planning and logic programming for dialog management. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 214–217. IEEE.

Kronlid, F. (2007). Implementing the information-state update approach to dialogue management in a slightly extended scxml. *Proceedings of the SEMDIAL*.

Larsson, S. and Traum, D. (2000). Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural language engineering*, 6(3&4):323–340.

Radomski, S., Schnelle-Walka, D., and Radeck-Arneth, S. (2013). A Prolog Datamodel for State Chart XML. *SIGdial Workshop on Discourse and Dialogue*. to appear.

Schnelle-Walka, D. and Radomski, S. (2012). A pattern language for dialogue management. In *Proceeding of VikingPLoP 2012*.

Schnelle-Walka, D. and Radomski, S. (2013). Probabilistic dialogue management. In *Proceeding of VikingPLoP 2013*. to appear.

Schnelle-Walka, D., Radomski, S., and Lange, A. (2013a). Voice-based error recovery strategies for pervasive environments. In *Proceedings of Spech in Mobile and Pervasive Environments, in conjunction with mobileHCI*. to appear.

Schnelle-Walka, D., Radomski, S., and Mühlhäuser, M. (2013b). Jvoicexml as a modality component in the w3c multimodal architecture. *Journal on Multimodal User Interfaces*, pages 1–12.

Williams, J. D. and Young, S. (2007). Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Biographical Sketch

Dirk Schnelle-Walka received his PhD from the Technische Universität Darmstadt in 2007. He transferred the idea of design patterns to the design of voice user interfaces. His research focus is on multimodal (collaborative) interaction in smart spaces. He is also the head behind several open source projects around speech technology, e.g., the open source voice browser JVoiceXML[1]. Dirk is a member of the ACM.

[1] http://jvoicexml.sourceforge.net

Niels Schütte

Dublin Institute of Technology
School of Computing
Dublin, Ireland

niels.schuette@gmail.com
www.comp.dit.ie/aigroup/?page_id=164

1 Research Interests

My research interest lies in the area of situated dialogue. In particular, I am interested in how dialogue is affected by the perception of the environment the participants of the dialogue have. At the moment, I focus on problems that arise because the dialogue participants have a diverging understanding of the environment, and on how these problems can be detected and resolved in a human-computer dialogue setting.

As part of my research I have previously looked at some aspects of reference in task based situated dialogues (Schütte et. al. 2010, ; Schütte et. al. 2011, ; Schütte et. al. 2012,).

I am currently putting together an evaluation scenario to compare different multimodal dialogue strategies in the context of a task based situated dialogue between a (simulated) robot and a human instruction giver. In the experiment, different problem conditions will be introduced into the robots perception, and the robot will then attempt to use different strategies to address the problems.

The following are some topics my research addresses:

1.1 Miscommunication about Perception

Visual perception by robot systems is potentially incorrect due to problems with object recognition and classification. If a robot misunderstands an object it perceives (e.g. if it mistakes a piano for a table), problems may occur if it has to discuss this object with a human dialogue partner. Ideally, a computer dialogue system needs to be able to detect such conditions, represent the fact that such conditions occurred and be able to pursue strategies to resolve the problem.

1.2 Common Ground in Situated Dialogue

To be able to talk about the environment, the dialogue partners need to form an agreement about what they perceive. I believe that this process can be understood as part of the more general grounding problem and be approached with similar strategies.

2 Future of Spoken Dialogue Research

I believe that an important thing dialogue systems research should address in the next years is to develop a strong example for a good practical application of dialogue systems. I think this will on one hand help to convince other researchers to appreciate the value of dialogue in human computer interaction, but also give important goals for the more practical development of dialogue systems.

I think that another important topic would be to generally address the problem of how to detect problems and how to gracefully deal with the fact that problem situations in a dialogue occurred.

3 Suggestions for Discussion

Some ideas for discussion:

- What are good examples of practical applications of dialogue systems?
- The use of different types of contexts into dialogue.
- What should dialogue systems do if dialogue breaks down?

References

Schütte, N., Kelleher, J., and Mac Namee, B. 2010. *Visual salience and reference resolution in situated dialogues: A corpus-based evaluation*. Dialogue with Robots. Papers from the AAAI Fall Symposium, Menlo Park, California. AAAI Press.

Schütte, N., Kelleher, J., and Mac Namee, B. 2011. *Automatic annotation of referring expression in situated dialogues*. International Journal of Computational Linguistics and Applications, 2(1 - 2).

Schütte, N., Kelleher, J., and Mac Namee, B. 2012. *A Corpus Based Dialogue Model for Grounding in Situated Dialogue*. Proceedings of the 1st Workshop on Machine Learning for Interactive Systems (MLIS-2012.)

Biographical Sketch

Niels Schütte is a PhD student at the Dublin Institute of Technology (DIT). He studied Informatics at the University of Bremen. He enjoys SCUBA diving and likes tea.

Stefan Ultes

Dialogue Systems Research Group
Ulm University
Albert-Einstein-Allee 43
89081 Ulm, Germany

stefan.ultes@uni-ulm.de
nt.uni-ulm.de/ultes

1 Research Interests

In the area of **Spoken Dialogue Systems** (SDS), my interest lies in rendering those systems more natural and user-aware. While current systems do not regard the current status of the system, using methods for **user state recognition** allows to enable **user adaptivity** in SDS. The general idea is to use **Machine Learning** techniques to automatically recognize the current user state to influence the decision process of the **Dialogue Management** module. There, an emphasis is placed on **Statistical Dialogue Modeling** in order to utilize a methodology which is able to deal with the uncertainty emerging from user state recognition inherently. To accomplish this, I am also interested in the fields of **User Simulation** and **Spoken Dialogue Assessment**.

1.1 Automatic User Satisfaction Recognition

For creating automatic recognition modules for user satisfaction (US), statistical models have to be trained. For supervised approaches, a target variable is required. It can either be acquired through system users or by expert raters annotating prerecorded dialogues. We addressed the question of which type of rating should be favoured in (Ultes et al., 2013) by analyzing a study containing both types of ratings with the outcome of recommending expert rater annotations.

For adapting the ongoing dialogue to US, the used paradigm has to fulfill specific requirements. In (Ultes et al., 2012b), we identified six items for a quality metric to be usable in adaptive dialogue management. The interaction quality (IQ) paradigm presented by Schmitt et al. (2011) based on expert annotations for US fulfills these requirements. It uses interaction parameters extracted from the dialogue modules which have been described thoroughly in (Schmitt et al., 2012).

For improving the recognition rate, we investigated several approaches. While Schmitt et al. used a static approach for estimating IQ, we investigated time-series models. While the utilization of a Hidden Markov Model (HMM) and a Conditioned HMM were not successful (Ultes et al., 2012a), applying a hybrid HMM with utilizing the confidence scores of static classifiers as observation probability showed better recognition rates. Furthermore, using a two-stage approach for IQ recognition using an error correction method also improved the recognition rate (Ultes and Minker, 2013b).

1.2 Quality-Adaptive Dialogue Management

The goal of my work is to automatically adapt the dialogue to the current IQ score. We presented two general ways of doing this: using rule-based dialogue management approaches and using statistical approaches (Ultes et al., 2012b). Rule-based adaption includes adapting the confirmation strategy or the initiative in the current dialogue situation. For adapting the dialogue to IQ using statistical approaches based on Partially Observable Markov Decision Processes, we will investigate adding the IQ value to the current system state as well as incorporating IQ into the reward function used for training.

We extended OwlSpeak, a spoken dialogue manager based on the Information State approach (Ultes and Minker, 2013a) by implementing the Hidden Information State approach by Young et al. (Young et al., 2007). The resulting dialogue manager facilitates both rule-based and statistical dialogue management utilizing the same dialogue descriptions which allows for easy evaluation of both types of adaptive dialogue. Furthermore, a quality recognition module has been added to OwlSpeak.

1.3 Future Work

While we already have addressed some aspects of user-adaptive spoken dialogue, still, many issues are to be tackled. For IQ recognition, there are other approaches which may be investigated. Conditioned Random Fields, which have shown to achieve good performance on other series labeling tasks, may result in better performance. Furthermore, IQ recognition has only been viewed as a recognition task for different classes. Regarding it as a regression task may also reveal new insight. Finally, as supervised approaches for IQ estimation require annotated data for training, unsupervised approaches might be interesting either for bootstrapping or for performing the actual estimation task.

For adaptive dialogue management, of course, the next step is to finalize the implementation of the statistical model. Then, different options for adapting the dialogue

to IQ can be investigated. Furthermore, rule-based adaption of the dialogue strategy to IQ has also be evaluated with extensive user studies.

2 Future of Spoken Dialog Research

In my opinion, one of the major aspects regarding the future of spoken dialogue research is to find means to render dialogue systems more naturally. And the key to this can be adopted from human-human-communication: humans do not solely rely on the linguistic content of speech. During the ongoing communication, humans perceive cues about the emotional state of the opposite person or can easily say if the opposite person is satisfied with the conversation, for example. All information about the user state is then intuitively used by us to influence what we say and how we say it. Furthermore, not only information about the user is of importance, but also information about the surroundings of the people talking to each other. The system has to be adaptive to the ever-changing environment just like humans are.

While there already are many research projects dealing with rendering the dialogue more user and situation adaptive, usually, the used approaches are only applicable for the specific problem the research is based on. However, having systems being adaptive to only one aspect is not sufficient. Therefore, I believe that we need to find unified models which allow to combine user and situation adaptive dialogue behaviour. Only by applying unified models and therefore being adaptive to multiple aspects at the same time will lead to more natural dialogues which also encompasses higher user acceptance.

3 Suggestions for Discussion

- Is user-adaptivity the right path to more natural dialogues and better user experience?
- How can we find unified models for user and situation adaptive dialogues?
- Is rule-based dialogue management still interesting for research?

References

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, Portland, Oregon, USA, June. Association for Computational Linguistics.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated corpus of the cmu let's go bus information system. In *International Conference on Language Resources and Evaluation (LREC)*.

Stefan Ultes and Wolfgang Minker. 2013a. Hisowlspeak: A model-driven dialogue manager with multiple control modes. In *9th International Conference on Intelligent Environments (IE 13)*, July. accepted for publication.

Stefan Ultes and Wolfgang Minker. 2013b. Improving interaction quality recognition using error correction. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, August. accepted for publication.

Stefan Ultes, Robert ElChabb, and Wolfgang Minker. 2012a. Application and evaluation of a conditioned hidden markov model for estimating interaction quality of spoken dialogue systems. In Joseph Mariani, Laurence Devillers, Martine Garnier-Rizet, and Sophie Rosset, editors, *Proceedings of the 4th International Workshop on Spoken Language Dialog System (IWSDS)*, pages 141–150. Springer, November.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2012b. Towards quality-adaptive spoken dialogue management. In *NAACL-HLT Workshop on Future directions and needs in the Spoken Dialog Community: Tools and Data (SDCTD 2012)*, pages 49–52, Montréal, Canada, June. Association for Computational Linguistics.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2013. On quality ratings for spoken dialogue systems – experts vs. users. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 569–578. Association for Computational Linguistics, June.

Steve J. Young, Jost Schatzmann, Karl Weilhammer, and Hui Ye. 2007. The hidden information state approach to dialog management. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–149. IEEE.

Biographical Sketch

Stefan Ultes studied Computer Science at the Karlsruhe Institute of Technology (Germany) with focus on "Cognitive Systems" and "Robotics". In 2010, he received his Diploma Degree in Computer Science (Dipl.-Inform.) when graduating on "Imitation Learning" in a joint project with the Carnegie Mellon University in Pittsburgh (USA). After working with SVOX Germany (now Nuance), he joined the Dialogue Systems Group at Ulm University under the supervision of Prof. Dr. Dr.-Ing. Wolfgang Minker in 2011 as a research assistant and PhD student. His topic is centered around user-aware dialogue rendering a human-computer-interaction adaptive to the perceived user state.

Martín F. Villalba

University of Potsdam
Karl-Liebknecht-Straße 24/25, Building 14
14476, Potsdam, Germany

martin.villalba@uni-potsdam.de
http://www.ling.uni-potsdam.de/~villalba

1 Research Interests

My main research publications are currently focused on **natural language interpretation**, and **situated language interpretation** in particular. Most of my research has been applied to **virtual environments**, using publicly available data from the GIVE Challenge as a basis from which to develop and test interpretation strategies. My current research interests are oriented towards two main branches: the importance of **user modeling** in interactive systems, and **natural human-computer interaction** and its place within dialog-intensive tasks.

1.1 Natural movement and behavior patterns in virtual worlds

In my undergrad work "Inference of Strategic Points in Virtual Worlds" (Benotti & Villalba, 2011) we have analyzed user behavior in the context of a task within a virtual world in order to look for patterns underlying typical user behavior. In order to achieve this, we designed a set of tasks in a 3D world, and tracked both the way in which players overcame obstacles and the routes they preferred to move around the world.

In this context, automatically locating conflictive points where the players were likely to require guidance proved to be straightforward. Furthermore, our approach also proved to be useful for determining strategically valuable locations within the task. Combining both sets of information, we can infer not only when to give feedback, but also how should an agent move from one point to another in such a way that mimics the navigation patterns of previous users. This allows us to design virtual agents displaying a ßofteränd more natural behavior when interacting with the user.

1.2 Instruction interpretation in virtual environments

I started my PhD studies focusing on the development of a system capable of giving instructions in virtual environments. By limiting our work to virtual environments, our work could then be ported to a wide range of applications and, in particular, it could aid the development of intelligent agents capable of guiding users through several tasks. As long as a user can be said to occupy a position and their actions have a reaction, our approach can be applied.

In this work, we developed a method for interpreting user instructions within a virtual environment, using data from the GIVE Challenge (Striegnitz et al., 2011) as test. We trained a system capable of following users' instructions without requiring a deep understanding of the structure of the given instruction. Instead, by applying lexical similarity metrics to compare new instructions with previously seen ones, we managed to obtain a high success rate in a fully automated system (Benotti et al., 2012). The approach leverages unannotated corpora and requires no manual annotation. A cornerstone of the success of our approach was taking the player's position as the first feature to analyze proved to be useful in reducing the search space for the correct interpretation of an instruction.

1.3 Resolution of referral expressions

In our latest work, we've focused on the problem of correcting misunderstandings of referring expressions, in the context of instructions given in virtual environments. For this project, we combined a *semantic* model with an *observational* model: the former would focus on the meaning of the referring expression, while the latter analyzes the user's behavior as part of a real-time grounding process. The probability distribution of each model was then combined to predict whether an instruction was correctly understood or not. In our first results the combined model has proven to be more accurate than either model by itself. The next planned steps for this line of research is the development of a full situated dialogue system, along with improvements in the overall accuracy results. A publication regarding these results is currently under review.

2 Future of Spoken Dialog Research

Dialog systems seem to be on an interesting spot right now. Today, the computing power required to bring spoken dialog systems to the public is at hand. High-speed Internet and cloud processing allows us to perform high performance tasks in portable devices, and most technological barriers have been now lifted. At the same time, the lowering of this barriers put in evidence the fact that a lot of the expectations in this area would eventually fall short. Voice recognition for home and office is per-

haps the most representative case: only after the technology was brought to the users, the community realized that keyboards where more efficient and convenient at the task, and also that users didn't really feel comfortable speaking by themselves.

Spoken dialog systems, then, are now in the ideal place: *true* dialog systems (namely, those in which the role of the speaker switches between the system and the user) can solve a certain family of problems better than any alternative, namely those tasks in which the path to the goal is not perfectly clear. Products such as Siri, Google Glass, virtual receptionists and GPS have taught us important lessons about what works, and perhaps most important, what doesn't. Many systems in many areas are now listening to us. It is the perfect moment, then, for systems capable of speaking back to enter the stage.

I feel confident that such a system will be implemented in the next few years. Metrics on the naturalness of dialog systems are still on the lower side of the spectrum, which I consider troubling if we intend to bring interactive dialog systems into the spotlight. Luckily, this metrics are improving, with some interesting strategies showing up in this direction.

3 Suggestions for Discussion

- Common ground in dialog systems: Can we use user and device metadata to improve our systems? Should we? And if so, which strategies have yielded the best results so far?
- Mobile dialog: Now that we can take our systems everywhere, which research areas are within our reach?
- The role of voice-enabled systems: Which is the one problem we *need* solved, but haven't managed to solve yet? Which one is our "voice-enabled killer app"? And if we can perform a task through dialog, does that mean that we *should*?

References

Luciana Benotti, Martín Villalba, Tessa A. Lau and Julián A. Cerruti. 2012. *Corpus-based Interpretation of Instructions in Virtual Environments*. The 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea.

Luciana Benotti and Martín Villalba. 2011. *Inference of strategic points in virtual worlds*. Proceedings of the 2nd Argentine Workshop on Videogames, Buenos Aires, Argentina.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller and Mariet Theune 2011. *Report on the Second Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5)*. Proceedings of the 13th European Workshop on Natural Language Generation (ENLG).

Biographical Sketch

Martín Villalba was born in Misiones, Argentina. He received his M.Sc. in Computer Science from the University of Córdoba. After a brief internship with IBM and some experience on System Administration, he is currently giving his first steps as a PhD student in the area of Computational Linguistics. He lives currently in Germany, where he's a student at the University of Potsdam.

His interest in fields outside Computer Science have given him an unusual look on his fields of study, as he is constantly trying to bring a touch of humanity into computer systems. He has also dabbled with electronics, game development, drawing and music.