ICT, Los Angeles, USA

Proceedings 2016

Workshop on Spoken Dialogue Systems for PhDs, PostDocs and New Researchers

September 16, 2016

Foreword

The Young Researchers Roundtable on Spoken Dialog Systems (YRRSDS) has been for many of us our first encounter with life in the research world. It aspires to be an environment in which young researchers can present late-breaking work, discuss ideas, connect with other researchers, and hear first person accounts of what it means to be a part of this field straight from respected senior members in the industry and academy.

There is an additional angle that is often overlooked, which is that YRRSDS is also the perfect stepping stone for anyone newly entering this field. Before the "big league" conferences, YRRSDS serves as an ideal forum for young researchers to discuss ideas in a laid-back setting with others who may have as many doubts and questions. This circulation of fresh perspectives has the potential for great impact on the next generation of research. The Roundtable's focus has always been to foster creative thinking and strengthen an international network of researchers. Past, present, and future YRRSDS participants form a pipeline of mentors and colleagues as former YRRSDS become senior professionals themselves. We are proud to be part of the chain of young researchers carrying these ideals into the present day.

We hope you enjoy the Roundtable. It's been a privilege for us to host its 12th edition, and it's our honor to have you join us. If you take away from YRRSDS as much as we have, we encourage you to volunteer for organizing next year's YRRSDS to keep the torch moving towards the future.

The YRRSDS 2016 committee

Organizing Committee

- Alexandros Papangelis, Toshiba CRL
- David Cohen, Carnegie Mellon University
- Eli Pincus, University of Southern California
- José David Lopes, KTH Royal Institute of Technology
- Ondřej Plátek, Charles University in Prague
- Lu Chen, Shanghai Jiao Tong University
- Maria Schmidt, KIT Karlsruhe Institute of Technology
- Ramesh Manuvinakurike, University of Southern California
- Tiancheng Zhao, Carnegie Mellon University
- Zhou Yu, Carnegie Mellon University

Advisory Committee

- Srinivas Bangalore, Interactions (former AT&T Research)
- Luciana Benotti, Universidad Nacional de Córdoba
- Rolf Carlson, KTH Royal Institute of Technology
- Maxine Eskenazi, Carnegie Mellon University
- Kallirroi Georgila, University of Southern California
- Jim Glass, Massachusetts Institute of Technology
- Kristiina Jokinen, University of Helsinki
- Tatsuya Kawahara, Kyoto University
- Diane Litman, University of Pittsburgh
- Wolfgang Minker, University of Ulm
- Sebastian Möller, Technische Universität Berlin
- Mikio Nakano, Honda Research Institute
- David Schlangen, Universität Bielefeld

- Gabriel Skantze, KTH Royal Institute of Technology
- David Traum, University of Southern California
- Nigel Ward, University of Texas at El Paso
- Jason Williams, Microsoft Research
- Steve Young, University of Cambridge
- Oliver Lemon, Heriot-Watt University
- Alex Lascarides, University of Edinburgh
- Sungjin Lee, Yahoo!
- Timothy Bickmore, Northeastern University
- L P Morency, Carnegie Mellon University
- David DeVault, USC Institute for Creative technologies
- Matthew Purver, Queen Mary University, London
- Julia Hirschberg, Columbia University
- Deepak Ramachandran, Nuance Communications
- Svetlana Stoyanchev, Interactions Corporations

Contents

1 Invited talks

1.1 Natural Language Generation for Dialogue Systems: What we (still) don't know. Marilyn Walker, UC Santa Cruz

Abstract: Dialogue interaction provides an extremely challenging and grounded evaluation of natural language understanding. Natural language understanding has greatly improved over the last 20 years, but what level of understanding is needed to respond appropriately in dialogue in a particular context? How can natural language generation produce these contextually appropriate responses? How can it scale? Should a conversational assistant have a persona or style? Should it entrain to the user, take the initiative, understand greetings and other phatic acts, and use an explicit model of dialogue structure? What are some near term and some long term solutions towards more natural dialogue interaction?

1.2 Stefan Scherer (USC ICT)

1.3 Gokhan Tur (Google Research)

2 Roundtables

The discussions are the core reason to held the workshop. They tend to be informal, futuristic and sometimes funny. The notes taken by the organizers is a subjective attempt to capture some conclusions.

2.1 Panel Discussion

Ethan: Most meaningful thing was internship. Working in industry and seeing that dreams of doing things which were industry experience. Everything doesn't materialize.

Traum: Everything doesn't materialize. Can you succeed as a student, researcher, find funding source etc. Any answers to those are meaningful.

Walker: Do something and knowing a lot more in next three months, I will know more. If it doesn't work, then it can be a baseline. So, working on things and making it better is the goal.

Trung: Learning many skills is important. Finding dialogue solutions on existing problems is hard. First success was applying a well know method to some other problem. That helps build credibility.

Gokhan: Cannot become a good researcher sitting at home. People were afraid to try new things. "Standing on shoulders of giants" is meaningful. Environment is meaningful.

How to deal with frustrations?

Walker: Dialogue picked me and I didn't pick dialogue. There was a point in time around 2000s when I thought that web access and smartphone was the best. I was worried that my research will all go waste. I didn't do something else. So, I had to be patient.

Ethan: Neural nets have been here for quite some time. But, got attention later.

Walker: Having patience is important. Its like an obsession. Its like doing art. Found analyzing human-human dialogue when SDS didn't work. So, I kept the motivation going. Started working with social media.

Gokhan: Don't follow the trend. Focus on the approach. Know that people are more interested in approach and the results. Results can be obtained with the approach. Build good approach.

Traum: Spoken vs text based dialogue. Text based was worked on in 80s. But, if ASR came in it would become irrelevant. Now we have a lot of things. Texting isn't going anywhere. It is not necessarily that it is important to pick and choose what I am really interested in. Pick the best tools for the task and don't pick the trend

How to implement dialogues into the real world?

Ethan: Its hard for people to understand the under the hood things. How to incorporate business rules into the things is important. Understand the customer. SDS are target understanding. The challenge is the wall garden. Users are exploring in the dark.

Can the users be told what to do?

Gokhan: Let users explore. Do more and more thing.

Truam: Go back to human conversation and see how to converse. You don't go to McD and ask about integral calculus. So, there is an environment and the reason the system

is built for. Give people a chance to explore the system and let them know what the system is built for. Draw on those models. How do you detect that the dialogue is not going anywhere and how to gracefully recover from there?

Walker: Human in the loop can work?

Ethan: Not ideal. It's a hard strategy to work around this problem. Humans could do better.

How would you hire someone to work with you?

Traum: Depends on the expertise from the people.

Gokhan: Deep learning

Traum: (Convert dialogue into a machine learning problem. Not good)

Walker: Broad coverage people are important. Inquisitive mind and someone who you can talk to about anything. Problem is that broad coverage has to be learnt from the right place. Finding those hard places is hard. Read a lot! Interest in sensitivity to language. Language is a way of communicating and not just a signal.

Ethan: needs a good judgement. I don't know is not an answer. Creating new things with all the recent tools is important. Look to learn as much as possible.

What is the ideal time and number of people?

Walker: 5-6.

Trung: 30 lot of different expertise.

Traum: Depends on the kind of dialogue system.

Traum: we're focused on language and a few people know a lot about language and more than us, but their contribution as a body of material to us is not quite useful. People from different discipline don't look at the problem the same way.

Walker: Lot of times people things everyone can do their own job. Miss out on lot of fun if you don't have technical skills if you can't do recent machine learn things. Ex: people can't contribute to the team if they're working on a different team. Its important that people are broad in their interest because they may fail to assimilate into different team culture.

What should deep learning people do to be accepted into dialogue community?

Ethan: give an example. Motivate well.

Marilyn: think hard about the baselines. Ask people about a good baseline. Ask people what side the issues are on. Have a good baseline for your tasks. Important to keep thinking about the problem and how you set it up.

Gokhan: happens in ML, my number is so much better is less influential. What was your contribution to make the model better ? How did your contributions affect the numbers ?

Trung: Deep learning cannot explain why it is better, and this is the reason why it works better and this is a good solution.

2.2 Roundtable: Commercial SDS

Two major categories in SDS:

- Eliminates labor: remove labor and provide utility. Replacing worker
- Ecosystem type: Siri etc. Makes apple's products more lucrative.
- Maluba: Agree with the notion.
- IOT provides lot number of use cases. Ex: ATT system, connected dialogue systems.
- Ondrej: May be it is not dialogue, may be integration system. Navigation system should ask questions.

Commercial dialogue systems are task oriented usually. AI in a game, playstation, talking to the AI and all work in a team. Entertainment system.

- Is siri a dialogue system ?
- Lacks states. May be it is a question answering system. Not a dialogue.
- But, it breaks the question down into multiple questions, so it is a dialogue system. QA system: rudimentary dialogue system. Proof by induction. So, it is a basic SDS. They have voice search. Maintain states and context.

Omilia: lot of customers. Conversational systems do not eliminate jobs. Goal is to not lose jobs but change jobs. Time is what matters. Orient agents to provide better customer experience.

Human in the loop AI can help leverage more complex stuff from humans.

Activation of card 2.5 euros cost if humans do the job. Agents can have a call and costs around 0.2 or 0.3. There is a major cost reduction.

Agents need to be able to grab attention. Lot of people just speak and are distracted. So, it doesn't quite grab attention.

Context of the dialogue: Ex: Hello .. customer missed it .. Hello again. Make the customers move on. If agent is confined to just finish conversation in a time bound manner, it cannot achieve the goal of optimality.

Customer service: dealing with a person is indivudial. But, the company cares about volumes serviced. How busy the queue.

Having a persona can help businesses perform better.

Imagine building a SDS for fridge, Task oriented system can be deployed without persona. It is like a siri use case.

People like talking to people. Siri: depends on use case.

Security systems should be smart: Robotic dog for security.

IOT: Fridge, camera and all the devices. Eco-systems creation with SDS.

People attribute personality to a voice. Even Roomba is associated a personality. Automatic systems is associated personality.

Alexa with hands free functionality is good and things don't quite need a personality yet.

Personality is probably needed not in the case where the functionality performs well but when it doesn't do well ? "I am sorry I didn't catch that" repeatedly saying is not a good solution and needs a personality.

How to recover from errors is important in commercial vs research sds ? Siri: default is google search.

In Europe customers have completely different attitude to the system.

As soon as they hear system they start swearing. Demographics might be important. Status of the customer is quite important to understand the logic. Learn from data coming in: Data and the systems need to be trained dynamically. The systems can learn from the data that they have already been input. How do these interactions adapt to the new kinds of data coming in?

2.3 Roundtable: Multi-modal Generation

Various channels that convey information How can we combine appropriate channels of information to convey multi-layered information (e.g. say something and indicate uncertainty via non-verbal signals) Feedback on generated output? Synchronization of output signals When do we actually need multiple modalities? How could MMG affect the way people communicate with the system?

Charts, graphs, other modalities besides just virtual agents; that is also MM generation Complementarity vs redundancy We can make some channels more concise if we use more (e.g. verbal & visual) Temporal sequences are different for different modalities but need to be coordinated somehow FML & BML for non verbal understanding and generation, NVBG, BEAT, CEREBELA Robots have other constraints, but VA may also be unnatural because they do not adhere to real laws of physics Output from various modalities may (collectively) be disruptive rather than helpful and we need to design methods of handling All these modalities (e.g. machines in a hospital - all are useful but together they are very noisy) Some modalities are more urgent / invasive than others (e.g. phone buzzing) What things matter in evaluating MMG? appropriateness, naturalness - different for different communities - perhaps measure how much people mimic the system For synchronization, we may have something that enforces rules, e.g. lipsync Most existing systems start with one channel and then augment it with other channels, rather than generating everything in parallel.

2.4 Roundtable: Multi-modal Understanding

- Various levels of understanding
- SLU, Non-verbal behaviour understanding
- How should we define the various levels of understanding?
- Can the various modalities be processed separately or not?
- Understanding based on long-term and short-term knowledge
- Synchronisation of signals
- Privacy issues - what kind of information do we keep / ignore?

Notes:

The fact we don't understand the French intonational patterns doesn't allow us to understand

Data is available, is little addressed in current dialogue systems

Needed for annotated data, lack of agreement between annotators

What multi-modal interaction: eye gaze in multi-party gazes.

Intonation as to complete the information of the asr.

Multi-modal understanding is dependent on the end goal.

Synchronize data: ping messages to synchronize time between machines. Latency is not often the same and currently is difficult to address in real time.

Feature representation is still a challenge.

Problem about synchronizing, there is a lot of data available but it is hard for processing the data. Lack of uniformity of in the way the data is collected. Take all the information that we want from one channel is hard. Timeline software that allows analysing all the streams: ELAN.

2.5 Roundtable: Evaluating Chatbots

crowd-source appropriateness of next response Stefan Ultes explored interaction quality Ryan - next utterance classification - is not perfect but better than Bleu David Traum coherence is not bad - quite correlate with user satisfaction permuted turn in dialogues multiple WoOz Turn by turn measures may be combined to global measure which is needed for chatbot Ask Tiancheng (Tony) DialPort evaluation platform (not only for chatbots) http://arxiv.org/abs/1606.02562 Marilyn suggest preparing shared-tasks for chatbot 10 years of Big Bang theory transcription as training data appropriateness vs. coherence appropriateness not much like dialogue Is very situation dependent Play it safe - is it desired? future expectation? do not repeat too often Summarize context desired especially if the user do not listen much propose next context We should focus on what user wants - task oriented vs chit-chat dialogue length is good measure

2.6 Roundtable: Statistical Method for SLU

- Has bunch of data labeled. Whats the best way to add new labels ? Slots and intent mapping. How addition of intents and slots? Needs recollecting and re-implementing the pipeline. May be generative models can be a solution to this. Needs re-training.

- Including context into the language understanding.

- Small labeled dataset, but large unlabeled dataset. Using self-taught learning.

- Using annotations on the ASR engine.

- DSTC2 data is static. Variability in the new data. Much more different from the dynamic data. So, SLU needs a service? Problem conversion between state tracking vs understanding.

- SLU: task oriented setting and slot fitting. Interested in non-task oriented setting? Understand what the person is speaking in open domain and respond differently. SLU task is much vague and depends on the domains.

- Generic understanding: AMR: Abstract meaning representation is more powerful than framenet.

- Spoken vs written: written is much harder because its harder to parse. But, spoken have shorter length and easier to parse.

- Give me a perfect AMR parser, but how to build the perfect SDS.

- Opentable API has three variable API. Query to the API which can fulfil the user request.

- Classification is over-simplified version of SLU.

- High precision grammar: disfluencies: Disfluency clean up doesn't clean up. Query rewrite is a big issue.

- Query rewrite is very dominant. Synonyms and disjunctions are added.

- Query rewrite: Speech is taken as a commodity.

- Incremental SLU could handle works by having barge-ins. Annoying but works. Query success tasks increased by performing incremental SLU.

- Retrain on real user utterances. Corpora is important.

- Residual nets are may be equivalent to RNN. Skipping layers and resnets don't quite have the same appeal due to depth and randomness.

- Wavenet: generates signals. Wavenets vs resnets ? Generating words is very different from generating images. Could still work. Recursive neural networks.

- Fit to the API is important. -Show me movies of "James Cameroon"- Build trees based on parse tree. Recursive neural networks needs a more structured training data.

- Recursive NN was more linguistically motivated.

- Bracketed representation and LSTM provides the structure.

- Memory augmented neural networks: Memory augments the learning by integrating context.

- Attention models vs memory networks.

- Memory networks are may be better suited for QA and not really dialogue.

- Direct linking works well for parsing lattices.

- Input layer is in the form a lattice. [wordlets]

- Memory networks [QA system] FB data set. Could be a good pointer. Memory network to represent the context.

3 Position papers

Ondřej Plátek

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské námestí 25
11800 Prague 1, Czech Republic

oplatek@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/ondrej-platek

1 Research Interests

My research focuses on automating task-oriented dialogue system deployment. In particular, I believe that end-to-end neural conversation agents will substantially simplify bootstrapping dialogue systems in narrow domain. Having a working dialogue system which can be optimized jointly using single objective function is crucial for my further research where I want focus on improving the dialogue system interactively from conversations.

The task of improving a dialogue systems over time has gained focus of many researchers. However, most of the works focus only on optimizing a dialogue manager with very limited action space (Gašić et al., 2011). Such system is able to optimize only action selection despite its most severe problem may lie in language understanding or natural language generation. Reinforcement learning of end-to-end system updates the system parameters so the most severe errors according to objective function are mitigated first.

The quality of task oriented systems depend not only how well they can communicate with user about their expertise - their domain, but to larger extent it depends also on quality of the database. A common problem even with high quality databases is that they easily obsoletes and the stored facts are no longer valid or incomplete. To mitigate the problem we would like to train a conversational agent to collect missing information from users.

Currently, I focus on simplifying the architecture of dialogue systems using neural networks models so one can train decent baseline from fewer data. So far, I have implemented dialogue-state tracking model (Plátek et.al, 2016) which uses encoder-decoder(Bahdanou et al., 2014) architecture which I would like to extend to end-to-end system. In recent work (**?**), a dataset from restaurant domain is collected with simple annotations which is intended to be used for training end-to-end system. Last but not least, I am interested in rapid prototyping of spoken dialogue systems using current technologies to test simple ideas e.g., how to extract information from users.

1.1 Work in progress: Task oriented End-to-End Dialogue Systems

I am currently working on a prototype of end-to-end dialogue system and its evaluation on Cambridge restaurant domain (Williams et al., 2013). We used the relevant part of the model to estimate the model potential on much easier task of dialogue state tracking (Plátek et.al, 2016) and verified that the model is able to capture dialogue history well. On the other hand, inspired from (Wen et al., 2016) we aim to optimize access to database jointly but without explicitly modeling the dialogue state. We plan to train the model with annotations of API calls instead of dialogue state labels. The work is still in progress, but we plan to evaluate the final model using coverage metrics, which prefer factual answers with similar entities based on the knowledge base ontology.

1.2 Future plans: Extracting annotations from users

In longer perspective, we aim at training the dialogue system so it will be able to collect explicit feedback - annotation interactively from users. We plan to explore how users react to incorrect replies. Later, we intend to detect a situations when user thinks he or she was provided with incorrect reply. Detecting misunderstanding is an important task since it is the first step to obtaining interactively a ground through annotation of the incorrect reply from users.

2 Future of Spoken Dialog Research

I suppose that in near future we will see massive use a very simplistic and narrow domain conversational agents through platforms like Wit.ai or Kick.com. The power of such platforms is the organic selection of domains which are convenient for such simplistic conversations. The obvious challenge is navigating the user to the desired narrow domain and both platforms do not use dialogue to do so effectively ignoring the problem. As as result multiparty dialogue will become important field.

In contrast the virtual assistants such as Siri, Cortana or Google Now will become more open-domain and more conversational agents. I assume that a lot of effort will

be invested in automatic scaling from a few dialogue domains to truly open-domain dialogue.[1] The neural networks model proved to scale well in terms of removing handcrafted components and modeling abstraction layers so the dialogue systems can be easily deployed. On the other hand, knowledge based approaches in question answering perform well and scale from simple domain to open-domain. From my perspective there are at at least two problems which need to be solved before these technologies can be successfully combined. First one need to solve how to parametrize the neural conversational models based on part of the knowledge graph under discussion. Second much more challenging task is to determine what is the knowledge graph under discussion using the neural model and the conversation history.

3 Suggestions for Discussion

My suggestions for the discussion roughly correspond to my research topics:

- End-to-End dialogue systems:
  - How to optimize database access jointly together with conversation?
  - What kind of additional supervision to provide to text-to-text task oriented conversational agents?
  - Is next utterance classification (Lowe et al., 2016) annotation feasible to obtain? Can it be used similarly as noise contrastive estimation (Gutmann, Michael and Hyvärinen, Aapo, 2010) for language modelling?
- Detecting misunderstanding:
  - How to detect misunderstanding automatically?
  - What training data is necessary for misunderstanding detection and how to collect them?
- Rapid prototyping of SDS:
  - Which technology to use?
  - Deployment strategies

References

[Povey et al. 2011] Povey, Daniel and Ghoshal, Arnab and Boulianne, Gilles and Burget, Lukáš and Glembek, Ondřej and Goel, Nagendra and Hannemann, Mirko and Motlíček, Petr and Qian, Yanmin and Schwarz, Petr and others 2011 *The Kaldi speech recognition toolkit*.

[Plátek et.al 2016] Ondˇrej Plátek, Petr Belohlávek, Vojt ˇ ech Hude ˇ cek, and Filip Jur ˇ cí ˇ cek Plátek, Ondřej and Bělohlávek, Petr and Hudeček, Vojtěch and Jurčíček, Filip 2016. *Recurrent Neural Networks for Dialogue State Tracking* To Appear in Slo-NLP 2016 Proceedings.

[Bahdanou et al. 2014] Dzmitry Bahdanau and Kyunghyun Cho and Yoshua Bengio 2014. *Neural Machine Translation by Jointly Learning to Align and Translate* CoRR, abs/1409.0473.

[Gašić et al. 2011] M Gašić, F Jurčíček, B Thomson, K Yu, S Young 2011. *On-line policy optimisation of spoken dialogue systems via live interaction with human subjects* Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop, 312-317.

[Williams et al. 2013] Williams, Jason and Raux, Antoine and Ramachandran, Deepak and Black, Alan 2013. *The dialog state tracking challenge* Proceedings of the SIGDIAL 2013 Conference, 404-413.

[Wen et al. 2016] Wen, Tsung-Hsien; Vandyke, David; Mrksic, Nikola; Gasic, Milica; Rojas-Barahona, Lina M.; Su, Pei-Hao; Ultes, Stefan; Young, Steve 2016. *A Network-based End-to-End Trainable Task-oriented Dialogue System* Arxiv, http://arxiv.org/abs/1604.04562.

[Lowe et al. 2016] Ryan Lowe and Iulian Vlad Serban and Michael Noseworthy and Laurent Charlin and Joelle Pineau 2016. *On the Evaluation of Dialogue Systems with Next Utterance Classification* Arxiv, https://arxiv.org/abs/1605.05414

[Gutmann, Michael and Hyvärinen, Aapo 2010] Gutmann, Michael and Hyvärinen, Aapo 2010. *Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.* AISTATS

Biographical Sketch

Ondřej Plátek is currently entering third second year as a PhD student in the area of Computational Linguistics at Charles University in Prague. Much of his experience with conversational agents comes from industry research focused not only spoken dialogue systems but also just on automatic speech recognition. He has enjoyed several freelancing opportunities for small companies. His interests include machine learning, linguistics, outdoor sports especially rock-climbing.

[1] I wonder if users truly want to know the information in the large knowledge graphs.

Ramesh Manuvinakurike

University of Southern California
Institute for Creative Technologies
Los Angeles, CA, USA

manuvinakurike@ict.usc.edu
http://manuvinakurike.com

1 Research Interests

My research interests lie in the area of **architecture of incremental spoken dialogue systems** (SDS), using **crowd-sourcing** to bootstrap SDS development and their **real world applications**.

1.1 Incremental spoken dialogue systems

Spoken dialogue systems are increasingly being used in applications. As the users get accustomed to such systems there is a need for building a more efficient and natural conversation system. Taking a leap from traditional non-incremental to incremental dialogue systems is not trivial. As the speech input from the users are continually evolving the systems need to update their understanding at a much faster rate and generate the responses naturally and as quickly as possible. In my research I am interested in exploring such architectures which make the dialogue systems more natural and hence more acceptable to users.

1.2 Crowd-sourcing to bootstrap SDS development

Crowd-sourcing provided plethora of opportunities to fast-track the process of SDS development. The popularity of crowd-sourcing for dialogue systems research has increased in recent times and has been quite extensively been used for collecting user interactions (manuvinakurike and DeVault 2015) (Zarrieß et.al 2015) and have been used successfully to build and deploy a spoken dialogue systems (Paetzel et.al 2015). Crowd-sourcing helps reduce time and cost for building and deploying a dialogue system compared to traditional in-lab methods (manuviankurike et.al 2015). Prototyping in quick time can help deploy and collect human-agent interaction data and further SDS development.

1.3 Real world applications

Support based systems, question answering and Personal assistants are the most widely used type of dialogue systems. However, as the SDS research progresses their potential in the field of education and medical interventions cannot be downplayed. Embodiment can play a major role and help develop effective interventions (manuvinakurike et.al 2014).

2 Future of Spoken Dialog Research

As SDS become a part of daily lives through personal assistants or automated tellers further avenues for their deployments will open up.

- Generates more interest in naturally conversing agents vs a voice based interfaces
- Generates more interest for better and domain independent language understand modules.

3 Suggestions for discussion

- Natural language understanding for reference resolution.
- Spoken dialogue system applications in industry.
- Incrementality in SDS research.
- Dialog and Question-Answering: Mutual benefits, user interaction.

References

Ramesh Manuvinakurike and David DeVault, Natural Language Dialog Systems and Intelligent Assistants, chapter Pair Me Up: A Web Framework for Crowd-Sourced Spoken Dialogue Collection, pp. 189– 201, 2015.

Sina Zarrieß, Julian Hough, Casey Kennington, Ramesh Manuvinakurike, David DeVault, and David Schlangen, “Pentoref: A corpus of spoken references in taskoriented dialogues.,” LREC, 2016

Maike Paetzel, Ramesh Manuvinakurike, and David DeVault, ““So, which one is it?” The effect of alternative incremental architectures in a high-performance gameplaying agent,” in SIGDIAL, 2015.

Ramesh Manuvinakurike, Maike Paetzel, and David DeVault, “Reducing the Cost of Dialogue System

Training and Evaluation with Online, Crowd-Sourced Dialogue Data Collection," in The 19th workshop on the semantics and pragmatics of dialogue, 2015.

Ramesh Manuvinakurike, Wayne F. Velicer, and Timothy W. Bickmore. "Automated indexing of Internet stories for health behavior change: weight loss attitude pilot study." *Journal of medical Internet research* 16.12 (2014).

Biographical Sketch

Ramesh Manuvinakurike just finished his 3rd year of his Ph.D program at University of Southern California. He is beign advised by Prof. David DeVault at USC. He has previously worked with Prof. Timothy Bickmore developing Embodied conversational agents for health interventions. He has worked with Prof. David Schlangen working on language understanding for reference resolution. He is interning at Adobe Inc. developing spoken dialogue system for image search and photo editing.

Felix Gervits

Tufts University
Human-Robot Interaction Lab
200 Boston Ave.
Medford, MA 02155

Felix.Gervits@tufts.edu

1 Research Interests

My primary research goal is applying models of human communication and dialogue to robotic systems in order to improve coordination in multi-agent human-robot teams..

One of the central challenges in this endeavor is the need to handle fast-paced natural language dialogue in performance settings. This is difficult because task-based dialogue is often very messy, riddled with disfluency, agrammatism, overlapping speech and ambiguity. Another challenge involves monitoring the mental state of one's teammates and performance of the team as a whole (Sycara and Sukthankar, 2006). The robot will need to monitor team communication channels to identify changes in performance state and team cohesion, and then modify its behavior to repair any problems.

There is a wealth of human factors literature on team performance that addresses how human teams overcome some of these challenges, and my research utilizes this knowledge base to inform the design of **dialogue systems for human-robot interaction (HRI)**. I also use methods from a variety of disciplines, including **corpus-based discourse analysis**, **team performance evaluation**, and **integrated architectures for NLU**.

1.1 Task-based dialogue in humans

My research has so far consisted of a systematic investigation of task-based dialogue in humans. The data was obtained from the Cooperative Remote Search Task (CReST) corpus (Eberhard et al., 2010), which contains hours of collaborative dialogue between remotely-communicating partners performing a joint-task. Since CReST was designed to simulate the structure of teams in which a robot may play a vital role (e.g., search and rescue missions), the results of this analysis inform our understanding of the kinds of communication and coordination strategies that artificial agents will need to adopt to be effective partners in these mixed-initiative teams.

To get at the linguistic- and dialogue-level properties in the corpus, I wrote a program to parse and filter out various features from the transcribed text data. These features include 5 types of disfluencies, pauses, speech rate, average utterance length, and dialogue moves. I then performed statistical analyses on these features to establish which if any were correlated with team effectiveness (Gervits et al., 2016).

Interestingly, I found that disfluency rate (particularly self-repairs) *increased* for the effective teams. This seemingly counterintuitive result suggests that disfluencies are not only caused by production difficulty, but rather can serve as collaborative tools in the discourse to enhance coordination and performance. Another novel finding showed that the best performing teams were those that more efficiently grounded their conversational exchanges and minimized joint collaborative effort through particular discourse patterns - namely: Check and Ready dialogue moves, frequent acknowledgments, and establishing shared referents.

These results have important implications for improving coordination in NLU systems and in mixed-initiative human-robot teams. The finding that self-repairs are strong indicators of collaborative process and are increasingly utilized in effective teams, suggests that the detection and identification of speech disfluency is crucial for robust NLU. For example, speech rate could signal increasing workload demands (Berthold and Jameson, 1999), self-repairs could indicate grounding, and the receptiveness of a speaker to monitor their teammate (Clark and Krych, 2004), and fillers can provide clues about turn-taking and discourse structure (Swerts, 1998). It is important that dialogue systems are able to utilize this information contained in disfluent utterances to gain insight into speakers' meta-cognitive states.

1.2 Integrated HRI architecture

The central focus of my research going forward involves implementing these markers of team effectiveness in the DIARC architecture (Schermerhorn et al., 2006). DIARC is an integrated cognitive-robotic architecture that also serves as a platform and test bed for HRI experimentation. It is a robust system which integrates vision, motor control, goal planning/action, as well as advanced natural language processing. On the language side, the system is capable of speech recognition, incremental parsing, reference resolution, and context-sensitive pragmatic reasoning.

My ongoing involvement with the architecture has

been to develop the dialogue manager in order to handle more interactive, dynamic exchanges - such as those found in the CReST corpus. I am working on several concurrent projects:

- Handling disfluent utterances. This presents a major challenge for current NLU systems, which generally cannot parse disfluent utterances. I am currently developing algorithms to detect different types of disfluencies, including self-repairs, filled pauses, and lexical fillers ("like", "you know"), and to enable the system to utilize information contained in the disfluencies to improve speech recognition.
- Contextual modulation to determine if the literal or nonliteral interpretation on an utterance is to be used. For example, the utterance "Can you walk forward?" could be interpreted literally as a question about the robot's capabilities, or nonliterally as an indirect command to walk forward. The goal of this project is to develop algorithms to automatically detect which meaning is intended based on contextual cues.

2 Future of Spoken Dialog Research

The ultimate goal of spoken dialogue research is to develop robust systems that can interact with humans in natural, real-world settings. Though we are currently nowhere near achieving this goal, I believe that we are furthest behind in the areas of 1) situated interaction, 2) extra-linguistic communication, and 3) disfluency handling. Given that human interaction largely takes place between embodied agents co-located in a shared physical environment, it is important that dialogue systems can make use of the shared perceptual context available in such settings. Additionally, dialogue systems should be able to handle facial expressions, gesture, back channel feedback, and many other non-verbal cues, as these are ubiquitous in natural human interactions. Finally, efforts at speech recognition and NLU need to be directed towards handling more natural kinds of utterances, which include disfluencies and agrammatical phrases of all sorts. Parsing should be incremental in order to respect human timing, and the system should have ways to handle errors and repairs in a seamless way in order to preserve the interaction if something goes wrong.

3 Suggestions for Discussion

- To what extent do design decisions for spoken dialogue systems need to be grounded in the empirical literature? What are the pros and cons of this approach?
- What are some ways to integrate non-verbal communicative signals from various modalities (facial expressions, gaze location, gestures, etc.) into the interpretation of an utterance? What are some existing systems that do this?
- How can "messy" features of speech (i.e., disfluencies) be detected and utilized by dialogue systems to improve performance.

References

Berthold, Andre, and Jameson, Anthony. 1999. *Interpreting Symptoms of Cognitive Load in Speech Input*. User Modeling: Proceedings of the Seventh International Conference, pp. 235-244.

Clark, Herbert H. and Krych, Meredyth A. 2004. *Speaking while monitoring addressees for understanding*. Journal of Memory and Language, 50(1), pp. 62-81.

Eberhard, Kathleen, Nicholson, Hannele, Keubler, Sandra, Gudersen, Susan, and Scheutz, Matthias. 2010. *The Indiana Cooperative Remote Search Task (CReST) Corpus.* Proceedings of the International Conference on Language Resource and Evaluation, (LREC) 2010, 17-23.

Gervits, Felix, Eberhard, Kathleen, and Scheutz, Matthias. 2016. *Team communication as a collaborative process*. Manuscript submitted for publication.

Schermerhorn, Paul, Kramer, James, Brick, Timothy, Anderson, David, Dingler, Aaron, and Scheutz, Matthias. 2006. *DIARC: A testbed for natural human-robot interactions*. Proceedings of AAAI 2006 Mobile Robot Workshop.

Swerts, Marc. 1998 *Filled pauses as markers of discourse structure*. Journal of Pragmatics, 30(4), pp. 485-496.

Sycara, K. and Sukthankar, G. 2006. *Literature review of teamwork models*. Technical report CMU-RI-TR-06-50.

Biographical Sketch

Felix is a PhD student at the Human-Robot Interaction Lab at Tufts University advised by Dr. Matthias Scheutz. He has a Bachelor's degree in Cognitive Science from Rensselaer Polytechnic Institute and a Master's degree in Linguistics and Cognitive Science from the University of Delaware. Felix has a diverse multidisciplinary research background, with experience in the fields of AI, Cognitive Neuroscience, and Psycholinguistics. He is currently using this diverse background to inform the design of robust, spoken dialogue systems for human-robot interaction.

Juliana Miehle

Institute of Communications Engineering
Dialogue Systems
Ulm University
Albert-Einstein-Allee 43
89081 Ulm, Germany

juliana.miehle@uni-ulm.de
http://nt.uni-ulm.de/miehle

1 Research Interests

In the area of **Spoken Dialogue Systems** (SDS), my research interest lies in rendering those systems more flexible while at the same time allowing a high degree of user-adaptiveness. While current systems are task-dependent or plan-based, I want to integrate ontologies which model domain information. Using ontology-based reasoning, the current domain state may be detected and forwarded to the **Dialogue Management** module. However, the decision process of the Dialogue Manager should not only rely on the domain information, but also take into account the user's current state including their culture and emotion. Therefore, two steps have to be performed in the Dialogue Manager: first, a state has to be modelled which takes into account all relevant information concerning the domain and the dialogue state and the user state. Second, an appropriate system action has to be generated. To accomplish this second part, I am interested in **Statistical Dialogue Modelling** and **Machine Learning** in order to utilize a methodology which is able to deal with the uncertainty emerging from the state modelling.

1.1 IQ-adaptive Statistical Dialogue Management using Gaussian Processes

In my Master Thesis, I examined the impact of the incorporation of the Interaction Quality metric (IQ) during the process of learning a policy for a Spoken Dialogue System. The Interaction Quality metric is thereby a measure for the user's satisfaction as described in (Schmitt and Ultes, 2015) which is automatically estimated at each dialogue turn. Since the trained policy decides which actions are taken next, the dialogue flow is adapted to the user's satisfaction. For the process of learning, Gaussian Process based Reinforcement Learning has been used as proposed in (Gašić and Young, 2014).

Afterwards, I investigated whether the incorporation of the IQ metric is beneficial. Therefore, different learning strategies with and without the IQ metric have been used to train different policies. Then, the performance of all trained policies has been evaluated regarding dialogue completion, task success, the average length of a dialogue and the average IQ value at the end of a dialogue.

In summary, one can conclude that the results concerning the incorporation of the Interaction Quality metric into the process of training a policy have great promise for the adaptation of the dialogue to the user's satisfaction and the use of the IQ metric as a measure for it. Both the use of the final IQ value at the end of a dialogue to decide whether the dialogue has been successful or not and the use of the current IQ value in every dialogue turn to calculate the immediate reward with which the policy is trained yield good results.

For future work on the subject of IQ-adaptive dialogue, the same adaptation techniques should be tested with real users as they might give new insight by showing unseen behaviour. Furthermore, other alternatives regarding the incorporation of the Interaction Quality metric may be applied and tested, e.g. the integration of the IQ in the belief space. It has to be analysed whether the IQ value may be used as a measure of the task success and which way is the best to incorporate it into the process of learning a policy.

1.2 Cultural Communication Idiosyncrasies in Human-Computer Interaction

In (Miehle et al., 2016), we investigated whether the cultural idiosyncrasies found in human-human interaction may be transferred to human-computer interaction. With the aim of designing a culture-sensitive Spoken Dialogue System, we designed a user study creating a dialogue in a domain that has the potential capacity to reveal cultural differences. The dialogue contained different options for the system output according to cultural differences. We conducted an on-line survey on the user's preference concerning the different options among Germans and Japanese to investigate whether the supposed differences may be applied in human-computer interaction.

Our results show that there are indeed differences, but not all results are consistent with the cultural models for human-human interaction. This suggests that the communication patterns are not only influenced by the culture, but also by the dialogue domain and the user emotion. Moreover, it is shown that not all cultural idiosyncrasies that occur in human-human interaction may be applied for human-computer interaction.

In this work, only one specific dialogue has been considered. To get a more general view and exclude effects which may depend rather on the domain than on the culture, in future work other dialogues from different domains should be examined. Furthermore, we have to identify how the defined cultural idiosyncrasies may be implemented in the Dialogue Management to design a culture-sensitive Spoken Dialogue System.

1.3 Future Work

While I made some first investigations in the two fields of Statistical Dialogue Management and Cultural Communication Idiosyncrasies in Human-Computer Interaction, I now want to use these results and address the Dialogue Management decision process which should take into account the user state as well as the dialogue state. For the dialogue state, I plan to integrate ontologies. My first aim is to model a state which takes into account all relevant information. Afterwards, I plan to use machine learning approaches to generate appropriate system actions.

2 Future of Spoken Dialog Research

In my opinion, one of the major aspects regarding the future research on Spoken Dialogue Systems is to find means to render the systems more flexible that one Spoken Dialogue System may be used for more domains and for larger domains without the need of being trained for every single sub-domain. Furthermore, the systems need to be developed further to companion systems, supporting their user with any task they have, thereby taking into account their current situation, their cultural and social background and their current emotional state. To accomplish this, we have to make use of paralinguistic information identified by psycholinguists that facilitates natural language understanding. The use of affective speech, gestures, body language, back-channel responses, and timing facilitates human-human communication and should therefore also be used for human-computer interaction. For dialogues to be successful between humans and machines, our Spoken Dialogue Systems have to be able to interpret all of these cues and react in an appropriate way.

3 Suggestions for Discussion

- Statistical Dialogue Modelling: Are pure Statistical Dialogue Modelling and Machine Learning methods really the answer?
- Reward Modelling: How do we tell the Spoken Dialogue System during training what is good or bad?
- Evaluation: Are there golden rules for the evaluation of adaptive Spoken Dialogue Systems? How do we measure the success of adaptation?

References

Juliana Miehle, Koichiro Yoshino, Louisa Pragst, Stefan Ultes, Satoshi Nakamura, and Wolfgang Minker. 2016. Cultural Communication Idiosyncrasies in Human-Computer Interaction. In *Proceedings of the SIGDIAL 2016 Conference*. Association for Computational Linguistics.

Milica Gašić and Steve Young. 2014. Gaussian Processes for POMDP-Based Dialogue Manager Optimization In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1):28-40.

Alexander Schmitt and Stefan Ultes. 2015. Interaction Quality: Assessing the Quality of Ongoing Spoken Dialog Interaction by Experts—And How It Relates to User Satisfaction. In *Speech Communication*, 74:12-36.

Biographical Sketch

Juliana studied Electrical Engineering at Ulm University (Germany) with concentration in ”Communication and System Technology”. In 2013, she received her Bachelor of Science (B. Sc.) and in 2015, she graduated on ”IQ-adaptive Statistical Dialogue Management using Gaussian Processes” and got her Master of Science (M. Sc.). After doing an internship at Nara University of Science and Technology (Japan), she joined the Dialogue Systems Group at Ulm University under the supervision of Prof. Dr. Dr.-Ing. Wolfgang Minker in 2016 as a research assistant and PhD student. Her topic is centred around user-adaptiveness and the integration of ontologies into Statistical Dialogue Management.

Zahra Razavi

Department of Computer Science
University of Rochester
14627, Rochester, NY

[srazavi@cs.rochester.edu]

1 Research Interests

My research interests lie in **Natural Language Processing** and its applications, in particular topics related to **dialogue systems**, **conversation planning**, and leading natural flowing conversation. In order to construct a human-like conversational agent, I also do some research on **knowledge representation and reasoning**, including the use of **common sense knowledge** and on-line **knowledge extraction** from the user's input throughout the conversation.

1.1 Current Research

I am currently working with several collaborators on developing a dialogue system called LISSA which is designed to help people improve their social skills (Ali et. al., 2015). LISSA is an avatar that tries to lead a natural conversation with a user, while at the same time providing some feedback on the user's non-verbal behavior. The particular challenge I am working on is to enable the system to hold a natural open-ended conversation is the challenge I am working on. In order to conduct the conversation, LISSA needs to 1) understand what the user says and 2) provide a reasonable, appropriate reaction to the user's utterance. This is done through a structure called a schema which contains the knowledge needed by LISSA to carry out a high-level task such as making a small talk about general topics. A schema consists of a dynamically modifiable outline of the events LISSA expects to occur in the dialogue, plus a hash-table which stores the interpreted information gained during the conversation.

The first version of LISSA with the automated dialogue system has been evaluated in a speed-dating task, in which the output transcripts were compared to the ones from the same tests with Wizard-of-Oz dialogue manager. The results show almost no differences between the rating results of transcripts from automated and Wizard-of-Oz techniques, judged by multiple research assistants who did not know the source of the transcripts.

We are also running a study on using a variant of LISSA for helping teenagers with Asperger Syndrome to improve their communication skills. The exploratory study showed success in terms of user satisfaction with the system (Razavi et. al., 2016). More studies are ongoing in order to confidently evaluate the system.

1.2 Future Works

The future goal for the dialogue system I am working on is to make the dialogue even more natural and effective. An outline of ways of accomplishing this is as follows:

1- Reasoning based on the knowledge extracted from the user's input: As we collect some information from the user's utterances and add it to our knowledge base, we are planning to enable the system to make inferences from the available knowledge and use the results later during the conversation. This should give the user a greater sense of naturalness and of being understood by the system.

2- In support of the above technique, we plan to insert some world knowledge, so that the inference system could make more human-like inferences throughout the conversation.

3- Another possible improvement of the system is to integrate the verbal and nonverbal behaviors of the system. Currently, the feedback on the user's nonverbal behavior is shown with some screen icons while the conversation is going on. However, the more natural way is to make the avatar comment on the user's nonverbal behavior at some appropriate points during the conversation.

4- It is also possible to have the avatar show appropriate non-verbal behaviors based on its understanding of the users input. These nonverbal behaviors include head nods, smiling and backchannel utterances.

5- Automated turn-taking is a challenging issue in current spoken dialogue systems. Many systems have the user control the turn-taking by pressing a button; we do the same in LISSA. However, in a natural dialogue system, the system needs to recognize the end of a turn and take its own turn automatically.

2 Future of Spoken Dialog Research

Dialogue systems have recently been of special interest both in academy and industry. The successful deployment of the early versions of personal assistants (like Siri, Cortana, etc.) encouraged research in the area. However, the existing spoken dialogue systems are still command based; the area of open conversation leaves much room for further exploration explore. I believe that a major focus in dialogue systems research in 5 to 10 years will be to make systems more interactive, human-like and able to have open-ended conversation. In order to reach that point, the system should have a reasonable plan for each conversation, extract knowledge from the user's input and have a reasoning ability that it can use along with various kinds of world knowledge, allowing it infer new information as humans do during a conversation.

Besides functioning as personal assistants, spoken dialogue systems could play other roles, including companionship (e.g. Smith et. al., 2011). Companionable systems have many potential applications, such as helping people improve their skills (Ali et al., 2015), tutoring children (Foster et. al., 2010), helping doctors to diagnose some psychological disorders (DeVault et. al., 2014), and helping lonely people such as elderly people to overcome their feeling of isolation.

Another area which I believe will make a great stride is the integration of language with other modalities. Implementing spoken dialogue in virtual agents needs to combine vision, language, and non-verbal behaviors of the agent in a natural and human-like way. Robots need even more complicated integration, since they should make meaningful head and body movements in 3D space during the conversation.

3 Suggestions for discussion

- How to design a flexible dialogue plan for an open-ended conversation?
- How to extract knowledge from the user's inputs during the conversation? How to represent the extracted knowledge so that it can be used for reasoning during the conversation?
- How can we insert world knowledge into system and make it available throughout the conversation?
- How could we launch human-like reasoning based on the knowledge extracted from user inputs and on stored world knowledge to show more genuine understanding?
- How could we integrate verbal and nonnonverbal behaviors in virtual agents and robots with an automatic dialogue manager?
- How could we automate the turn-taking task in a conversation with a human, so that the machine could confidently recognize the end of turn and take the turn?

References

S. Zahra Razavi, M. Rafayet Ali, Tristram H. Smith, M. Ehsan Hoque and Lenhart L. Schubert. 2016. *Can chatting with a virtual human help Aspies improve conversational skills? An initial exploration*, Proceedings of the 16th International Conference on Intelligent Virtual Agents, LA, US (in press)

M. Rafayet Ali, Dev Crasta, Li Jin, et al. 2015. *LISSA- Live Interactive Social Skills Assistance*, In International Conference on Affective Computing and Intelligent Interaction (ACII) IEEE. 173-179.

David DeVault, Ron Artstein, Grace Benn, et al. 2014. *SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support.* International Conference on Autonomous Agents and Multi-Agent Systems, 1061–1068.

Cameron Smith, Nigel Crook, Johan Boye, et al. 2011. *Interaction strategies for an affective conversational agent*, In International Conference on Intelligent Virtual Agents, Springer Berlin Heidelberg, 301-314.

Mary E. Foster, Katerina Avramides, Sara Bernardini, et al. 2010. *Supporting children's social communication skills through interactive narratives with virtual characters.* In Proceedings of the 18th ACM international conference on Multimedia, 1111-1114.

Biographical Sketch

Zahra Razavi is pursuing her PhD in Computer Science department at university of Rochester. As a second year PhD student, she is working in the Natural Language group under supervision of Professor Lenhart Schubert. She earned her Bachelor and Master in Electrical Engineering from Sharif University of Technology and University of Tehran respectively.

Ondřej Dušek

Charles University in Prague
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25
11800 Prague 1, Czech Republic

odusek@ufal.mff.cuni.cz
http://ufal.mff.cuni.cz/ondrej-dusek

1 Research Interests

My main research focus is **natural language generation** (NLG) for spoken dialogue systems (SDS). I have been working with template-based and rule-based NLG systems, and I am developing a fully trainable NLG system that is based on **sequence-to-sequence learning** using recurrent neural networks. Recently, I have been focusing on making the NLG outputs **user-adaptive**. I also concern myself with NLG for languages with rich morphology (which include Czech, my native language) and **multilingual** natural language processing in general.

I am also interested in machine translation, which often uses techniques similar to NLG. Since I have a backrgound with a strong emphasis on linguistic analysis, my other research interests include dependency syntax, deep syntax (within the Prague tectogrammatics theory (Sgall et al., 1986)), and valency.

1.1 Previous Work

My most recent work is an NLG system trainable using just pairs of input meaning representations (such as dialogue acts) and the corresponding output sentences.[1] The first generator version was a two-stage setup composed of a sentence planning step based on A*-style search with a perceptron ranker and a rule-based surface realizer (Dušek and Jurčíček, 2015). The second version is based on sequence-to-sequence models and is able to operate in a two-stage setup, generating sentence plans for the surface realizer, as well as in a joint setup generating strings directly (Dušek and Jurčíček, 2016b). My recent experiments focus on entrainment, i.e., adapting the generator output to previous user utterances by reusing wording and syntax (Dušek and Jurčíček, 2016a).

My other work involves various improvements to the TectoMT transfer-based deep syntactic machine translation system, which includes a natural language generation component (Žabokrtský et al., 2008; Dušek et al., 2012; Dušek et al., 2015).

I have also worked on domain expansion, database interface, handcrafted dialogue policy, and the NLG component of the Alex SDS in the public transport information domain (Dušek et al., 2014).[2] Even though this work mostly involved software development of standard SDS components, it helped me understand the inner workings of a dialogue system.

Previously, I have developed a trainable morphology generation system based on logistic regression, word suffix features, and edit scripts that convert base word forms into correctly inflected forms (Dušek and Jurčíček, 2013).[3]

1.2 Future Plans

I am currently finishing my Ph.D. thesis, which is concerned with improvements of NLG for dialogue systems. My immediate goals therefore involve improvements to my sequence-to-sequence NLG system and experimenting with **generating Czech** sentences, where the rich morphology introduces different kinds of problems than those that need to be addressed in English.

I am also interested in a tighter integration of the individual models in the traditional dialogue systems pipeline, possibly resulting in **end-to-end** fully trainable solutions (Wen et al., 2016). Within this scenario, I would like to focus more on custom-tailoring the system responses for a particular user and situation – adapting to users' way of speaking and providing contextually accurate responses.

2 Future of Spoken Dialog Research

I think that the most important problems in SDS to be solved in the near future are the ones concerning fast deployment for multiple and/or open domains. Current large-scale virtual asistants developed by the industry, such as Siri, Cortana, or Google Now as well as new chatbot platforms such as Facebook's Viv offer a wide variety of domains but their inner workings are handcrafted in a large part, and their ability to keep track of dialogue contexts is rather limited.

A progress beyond the current systems will require improved architectures and training methods, where a turn

[1] https://github.com/UFAL-DSG/tgen

[2] https://github.com/UFAL-DSG/alex

[3] https://ufal.mff.cuni.cz/flect/

towards end-to-end training (Wen et al., 2016; Williams and Zweig, 2016) and bootstrapping with little training data are to be expected.

In addition, more expressive semantic representations of the dialogue state will be required for open domains and more flexibility. Here, I can imagine a shift towards graph-based knowledge representation and later possibly towards implicit modelling of the dialogue state, as is the case in today's experimental chatbots (Li et al., 2016).

Another future research direction in dialogue systems leads towards more adaptivity to a particular user and producing a more natural impression by eliciting richer contexts, conforming to social conventions, and adapting to the user's way of speaking.

Futher, future dialogue systems will probably support multimodal interfaces at a broader scale than today, combining touch input and visual output with the speech interface. Finally, I expect better support for multiple languages in SDS.

3 Suggestions for Discussion

My suggestions for the discussion roughly correspond to my research topics:

- NLG in dialogue systems – how to achieve fluency, naturalness, and responsiveness at the same time? How to control the output quality of statistical NLG?
- End-to-end task oriented systems and chatbots – can the latter inspire the former?
- Open-domain dialogue systems – how to make the dialogue systems scalable?
- Multimodality in dialogue systems – how to make use of multimodal interfaces while preserving a natural dialogue?

References

O. Dušek and F. Jurčíček. 2013. Robust multilingual statistical morphological generation models. In *Proceedings of the ACL Student Research Workshop*, page 158–164, Sofia, Bulgaria.

O. Dušek and F. Jurčíček. 2015. Training a natural language generator from unaligned data. In *Proceedings of ACL*, Beijing, China.

O. Dušek and F. Jurčíček. 2016a. A context-aware natural language generator for dialogue systems. In *Proceedings of SIGDIAL*, Los Angeles, CA, USA. To appear.

O. Dušek and F. Jurčíček. 2016b. Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany. To appear.

O. Dušek, Z. Žabokrtský, M. Popel, M. Majliš, M. Novák, and D. Mareček. 2012. Formemes in English-Czech deep syntactic MT. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, page 267–274, Montreal.

O. Dušek, O. Plátek, L. Žilka, and F. Jurčíček. 2014. Alex: Bootstraping a spoken dialogue system for a new domain by real users. In *Proceedings of SIGDIAL*, pages 79–83, Philadelphia.

O. Dušek, L. Gomes, M. Novák, M. Popel, and R. Rosa. 2015. New Language Pairs in TectoMT. In *Proceedings of the 10th Workshop on Machine Translation*, pages 98–104, Lisbon, Portugal.

J. Li, W. Monroe, A. Ritter, and D. Jurafsky. 2016. Deep Reinforcement Learning for Dialogue Generation. *arXiv:1606.01541*.

P. Sgall, E. Hajičová, and J. Panevová. 1986. *The meaning of the sentence in its semantic and pragmatic aspects*. D. Reidel, Dordrecht.

T.-H. Wen, M. Gašić, N. Mrkšić, L. M. Rojas-Barahona, P.-H Su, S. Ultes, D. Vandyke, and S. Young. 2016. A Network-based End-to-End Trainable Task-oriented Dialogue System. *arXiv:1604.04562*.

J. D. Williams and G. Zweig. 2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. *arXiv:1606.01269*.

Z. Žabokrtský, J. Ptáček, and P. Pajas. 2008. TectoMT: highly modular MT system with tectogrammatics used as transfer layer. In *Proceedings of WMT*, page 167–170, Columbus, OH, USA.

Biographical Sketch

Ondřej Dušek is a research assistant and a 4th-year Ph.D. student of computational linguistics at the Institute of Formal and Applied linguistics, Charles University in Prague, focusing on natural language generation in spoken dialogue systems and structural machine translation. His previous studies provided him with a background both in computer science and theoretical linguistics: he obtained his master's degree in computer science in 2010 and a master's degree in German philology in 2013, both at Charles University.

His academic experience includes work on the European Union projects FAUST (2011-2013) and QTLEAP (since 2013), both on machine translation, and the VYSTADIAL project of the Czech Education Ministry focusing on spoken dialogue systems (since 2012).

His non-work interests include music, movies, a bit of sports, and spending time with friends.

[Eugenia Anya Hee]

[University of Southern California]
[Institute for Creative Technologies]
[12015 E Waterfront Dr.]
[Los Angeles, CA 90094]

[ehee@ict.usc.edu]

1 Research Interests

My research interests lie generally in the area of **human computer interaction** and the ways in which language can create **multimodal grounding** between human participants and virtual agents. Grounding is defined generally as "the augmentation of common ground as the production of contributions" (Traum, 1999) between any or all participants involved. Learning what causes this development of common ground is essential in spoken dialogue systems because by studying the human ability to communicate, we can better understand the ways in which we interact with machines across variety of different scenarios. Furthermore, we can study what builds trust and connections between humans and apply these same techniques to automated systems to produce more helpful and efficient machines. I have previously studied human interaction and language, particularly **word choice,** between people in the field of psychology and linguistics, but I have also studied human dialogue between an embodied virtual agent, a disembodied virtual agent, and an embodied robot. Much of my work includes **analyzing data** from these recordings to annotate and code for signs of mutual grounding or cases of non-understanding, both **verbal and non-verbal**. By studying human language in as many varied scenarios as possible, I hope that I can eventually work with spoken dialogue systems in the future and develop trust and mutual understanding between humans and their technology.

1.1 Past Research

As part of the USC Family Studies Project Lab last year, I worked extensively with the Home Data team and used the Audacity software to help transcribe and analyze hundreds of recordings of couples throughout their daily lives. The purpose of this study was to find a correlation between couple's overall relationship satisfaction and their everyday dialogue. In particular, we focused on word choice, and the phenomenon that I helped study was the use of first-person pronouns in everyday conversation, (e.g. "I" "I'm", and "I'll"). Our prediction was that high first person pronoun use would lead to less satisfaction between partners. For women, this was true overall, but in times of conflict, higher first person pronoun use actually led to more relationship satisfaction (Power et al., 2016). These were important findings because they showed a possible correlation between language and satisfaction, implying trust and overall happiness. When dealing with spoken dialogue systems, these findings could also be used as a guide to avoid first person pronoun use in most conversational language, but to use first person pronouns more during times of conflict, to take the blame for mistakes or to show support and understanding (Ex: "I know, I was wrong").

1.2 Present Research

Currently, I am working at the Institute for Creative Technologies, studying human interactions with various different virtual characters, both embodied and disembodied. As part of the New Dimensions in Testimony project, we are working with pre-recorded clips of a Holocaust survivor, where participants can ask questions and receive real time responses through voice recognition and an automated system that chooses an appropriate answer (Traum et al., 2015). By having this form of active, spoken dialogue between the participant and the survivor, we hope to instill passion and interest in their stories in a way that could not be accomplished with regular video. The way that these participants have interacted with the survivor sheds light onto a new form of interactive education that can promote interest and withstand the test of time, and I see many potential applications in the future with this technology. Additionally, I am also working on a Robot Grounding project that analyzes signs of grounding in a participant as he or she interacts with a disembodied virtual human and an embodied robot as they navigate through two different ranking tasks. We also study the effects of having the participant talk about favorite activities and personal information to see if this would develop a stronger bond between the agent and the participant. Data collection is still currently ongoing, and I am helping with annotations of different signs of grounding.

1.3 Future Research

In the future, I hope to continue in the field of spoken dialogue systems and natural language processing to improve the current systems and develop context dependent responses as well as more robust voice recognition systems. I am curious to see if spoken dialogue can be used in more varied environments, like inside a classroom with young students as a form of education, or inside nursing homes with the elderly as a form of companionship. With the way technology is advancing, I see applications beyond just entertainment, and I hope that my research in the future can help me discover these new ways to use technology to provide aid.

2 Future of Spoken Dialog Research

In my opinion, I believe that spoken dialogue research will be used in the future to benefit people across a wider range of fields. Currently, we have systems available on personal laptops and mobile devices, but I anticipate that in the next five to ten years, this will grow to reach classrooms and larger institutions, providing aid in the form of education or therapy. I believe this generation of young researchers has no limit to the kinds of discoveries they will make, and I hope that we will be able to discover the technology needed to develop spoken dialog systems on a much larger scale. The kinds of questions to investigate would be ones regarding error checking and scalability, as well as understanding how to enhance overall user experience rather than detract from it. Questions like "How can we build a more robust natural language processing system built without context dependency so that we can facilitate more natural conversation?" and "How can we create a system that can be understood and appreciated by the youth or the elderly?" come to mind when thinking of research in the future.

3 Suggestions for Discussion

I would be interested in any discussion regarding spoken dialogue systems, but in particular these few topics:

- Ethical and cultural standards when dealing with research in human computer interaction and societal stereotypes of artificial intelligence
- Context dependency in dialogue and potential ideas to deal with context dependent questions and answers
- More frequent connectivity between researchers whether it be through conferences or a technological interface
- Examining user needs and working with marketing specialists who can understand the benefits of spoken dialogue research and advise where it can best be applied

References

Clark, H. H., and Brennan, S. E. 1991. Grounding in communication. In Resnick, L. B.; Levine, J.; and Teasley, S. D., eds., Perspectives on Socially Shared Cognition. APA.

Kaitlyn A. Power, Sohyun C. Han, Adela C. Timmons, Laura Perrone, Jamie Nguyen, & Gayla Margolin. May 2016. *"I" usage in couples' everyday lives: How language relates to relationship functioning.*

David R. Traum. Computational Models of Grounding in Collaborative Systems, in working notes of AAAI Fall Symposium on Psychological Models of Communication, p. 124-131, November, 1999.

David Traum, Andrew Jones, Kia Hayes, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. New Dimensions in Testimony: Digitally Preserving a Holocaust Survivor's Interactive Storytelling. In *ICIDS*, Cophenhagen, Denmark. *Lecture Notes in Computer Science*, Vol. 9445, pp. 269-281, Springer International Publishing Switzerland, 2015.

Biographical Sketch

Eugenia (Anya) Hee is an undergraduate student at the University of Southern California studying computer science and cognitive science. She hopes to pursue higher level education in computer science with an emphasis in artifical intelligence. She has experience as a research assistant both at the University of Southern California as well as the Institute for Creative Technologies affiliated with the university. In her free time, she enjoys listening to music, traveling, and baking desserts.

Yanchao Yu

Heriot-Watt University
School of Mathematics and Computer Science
Edinburgh Campus
Edinburgh EH14 4AS
United Kingdom

y.yu@hw.ac.uk
sites.google.com/site/yanchaoyu147/

1 Research Interests

My primary area of research interest lies in **Perceptual Semantic Grounding** using **Natural Language Interaction**. More specifically, I am working on building a multimodal teachable system that can interactively learn to ground, identify, and talk about objects and low-level features (visual attributes) in the real environment through dialogue from the human tutors. The system allows learning novel knowledge abut objects from scratch based on semantic and perceptual processing in the dialogue.

I am interested in exploring what linguistic approaches and dialogue strategies may help to improve the system/robot's learning performance, and also encourage humans to engage in a life-long learning game/task. It may ranges over different fields of **Incremental Dialogue Management**, **Semantic Analysis of Perceptual Scenes** as well as **Continuous Perception Learning**. On the current research stage, I am working on a fully adaptive learner dialogue strategy using Reinforcement Learning (RL) to achieve a best trade-off between learning performance and the effort by human tutors.

1.1 Semantic Language Grounding

There are two levels of language grounding that a multimodal system should consider when learning from human tutors through dialogue:

1) *Dialogue Level*: language grounding at the dialogue level is related to dialogue context from all participants and corresponding contextual analysis. It reflects whether the previous context has been agreed by participants or not and who has agreed with it.

2) *Perceptual Level*: language grounding at the perceptual level takes information from different modalities into account. It focuses on aligning words/phrases/sentences in NL to a perceptual scene (e.g. images and events) in the physical world.

In terms of Human-Robot Interaction and collaboration, an intelligent system/robot needs to consider both levels as it expects to interact with humans in a more natural human-like way within a real environment.

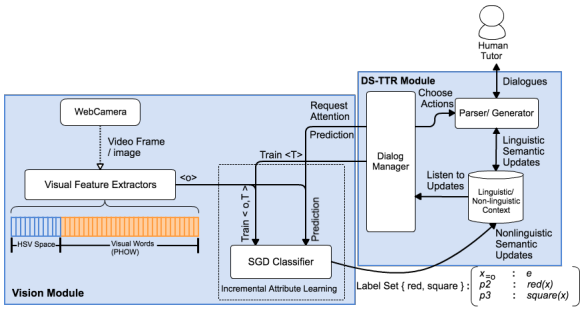


Figure 1: Architecture of the Teachable system

1.2 Previous Work

My current research mainly focuses on language grounding at the perceptual level. I will briefly describe my work in this section.

1.2.1 Integrating Semantic Processing with Vision

Yu et al. (2015) developed a multimodal teachable system for interactively learning to identify novel visual attributes through dialogue from human tutors. This system consists of two key components – a *vision system* and the *DS-TTR parser/generator* (see Fig. 1). The vision module classifies a (visual) scene, i.e. deems it to be of a particular type, expressed as a TTR Record Type (RT). This is done by deploying a set of binary and incremental attribute classifiers (Logistic Regression SVMs with Stochastic Gradient Descent (see Yu et al. (2016a)), which ground the simple types (atoms) in the system (e.g. 'red', 'square'), and composing their output to construct the total type of the visual scene. This representation then acts as a) the non-linguistic context of the dialogue for DS-TTR and b) the logical database from which answers to questions about object attributes are generated. The system can also generate questions to the tutor (Yu et al., 2016b) about the attributes of objects based on the entropy of the classifiers that ground the semantic concepts, e.g. those for colour and shape. The tutor's answer then acts as a training instance for the classifiers (basic, atomic types) involved.

1.2.2 Investigating Dialogue Strategies for Interactive Learning

Yu et al. (2016a, 2016b) have investigated the effectiveness of different tutor/learner-driven dialogue policies and capabilities (e.g. initiative, uncertainty, context-dependency as well as knowledge-demanding) in an interactive learning process with humans. We intend to explore an appropriate policy with best trade-offs between learning performance (i.e. recognition accuracy) and effort required by the tutor. Through comparisons, a dialogue strategy that the system takes initiative and consider uncertainty from the visual classifiers has demonstrated a better performance than the others.

1.3 Current & Future Works

Currently, we are focusing on optimising the Learner dialogue strategy with an adaptive uncertainty threshold using RL (Yu et al., 2016c). The partially adaptive dialogue strategy may determine when and how to express novel visual objects or ask for help from human tutors based on its previous knowledge and experience by adapting the uncertainty threshold.

In the future work, we intend to 1) work on a data-driven, incremental dialogue management at the lexical level; 2) use the similar setup to collect multimodal data (language and vision examples) from a set of human-human game interactions. The goal of these works is to explore and learn a human-like dialogue policy that may handle the interactive learning problem with humans in a more natural way.

2 Future of Spoken Dialog Research

In my opinion, there are many possibilities for spoken dialogue systems to keep growing in the near future, e.g. open-domain, data-driven dialogue management and incremental processing in dialogue system. I expect a continued growth in this field related to grounding problems at both dialogue and perceptual levels. An effective spoken dialog system need to be able to not only understand what users are talking about, but also ground situated context into other modalities from the real environment. To achieve this goal requires more close collaborations between researchers from different fields, such as physics, computer vision, multimodal interaction, and robotics.

In the more distant future, dialogue systems will be deeply involved in Human-Robot Collaboration, instead of the standard interaction. Apart from grounding, the important aspects of these collaborations might also include identifying and distinguishing user preferences and also detecting the final purpose through dialogue. A unified dialogue framework needs to be proposed to be deployed for different collaboration tasks and then expanding and sharing experience from one task to another.

3 Suggestions for Discussion

These are the four topics I propose for discussion:

- The new position of spoken dialogue systems in the future technology with new mediums (e.g. the 3D virtual reality);
- The importance of learning perceptual information online through Dialogue with humans.
- The possibility of training a dialogue manager with data from multiple modalities, rather than only speech examples.

References

Yanchao Yu, Arash Eshghi and Oliver Lemon 2016. *Comparing attribute classifiers for interactive language grounding*. The workshop on vision and language (VL15) associated with EMNLP. Lisbon, Portugal.

Yanchao Yu, Arash Eshghi and Oliver Lemon 2016. *Training an adaptive dialogue policy for interactive learning of visually grounded word meanings*. The Seventeenth Annual SIGDial Meeting on Discourse and Dialogue (SIGDial). Los Angeles, USA.

Yanchao Yu, Arash Eshghi and Oliver Lemon 2016. *Interactive Learning of Visually Grounded Word Meanings from a Human Tutor*. Fifth ACL workshop on Vision and Language (VL'16). Berlin, Germany.

Yanchao Yu, Arash Eshghi and Oliver Lemon 2016. *Comparing dialogue strategies for learning grounded language from human tutors*. Twentieth Workshop on Semantics and Pragmatics of Dialogue (SEMDIAL). New Jersey, USA.

Biographical Sketch

Yanchao Yu is a Research Assistant and part-time PhD student under the supervision of Professor Oliver Lemon in the Interaction Lab at Heriot-Watt University in Edinburgh, Scotland. He received his Master's degree in Software Engineering from Heriot-Watt University. During his Masters studies, his first encounter with spoken Dialog system was his Master project that built a mobile intelligent assistant for emergency aid. After that, he worked on several mobile spoken dialogue systems (e.g. PARLANCE and SpeechCity) in the Interaction Lab and started his part-time PhD in the lab in October 2014. Other than research, he is also interested in entrepreneurship as a career plan in the future.

Xiaoyun Wang

Doshisha University
Department of Information and
Computer Science
1-3 Tatara Miyakodani, Kyotanabe-shi,
Kyoto, JAPAN 610-0394

ou.gyouun@gmail.com

1 Research Interests

My research interests lie generally in the area of **spoken dialogue systems** aim at **foreigner languages learning**. The main theme of my research is exploring the effective and customized **acoustic modeling** for the **non-native speech** used in a dialogue-based computer-assisted language learning (CALL) system. These kinds of dialogue systems act as automated interlocutors that prompt learners to elicit speech in the target language and provide informative feedback that is of enormous educational value in terms of improving the learners' language communication skills (Kawai, 1997; Ito, 2008). But somehow, the quality of non-native speech usually differs significantly from native one in terms of phonemes, prosody, lexicon, disfluencies, and so on. Its automatic speech recognition (ASR) can be a great challenge. My research goal is to build a recognition system with adapt models that suit the character of second language (L2) speakers.

1.1 Dialogue-based computer-assisted language learning (CALL) system

I am interested in computational linguistics, and modeling based on the characteristics of both acoustic and linguistic features for speech, such as language acquisition based on human-computer/robot interaction; such as phonetic acoustic features of speech, and lexical or grammatical information of transcription. In addition I design a dialogue-based English CALL system for Japanese, and modeling the customized acoustic models for Japanese-English speakers.

Wang et al. (2014) described a work on developing a human-machine dialogue-based English CALL system. The system aims at eliciting more speech production from Japanese learners in order to improve their speaking skills, by involving them in man-machine dialogues. The system interface provides prompt text in Japanese, in order to help learners construct utterances which are easy to be recognized. Figure 1 illustrates a screenshot of the man-machine interface of the developed CALL system.

Figure 1: *A screenshot of the CALL interface: (1) dialogue scenario selection (shopping, restaurant and hotel); (2) prompt by system; (3) hint stimulus; (4) recognition result; (5) corrected feedback*

1.2 Customized acoustic modeling with reduced phoneme set considering integrated acoustic and linguistic features of second language speech

Most of the ASR technologies have been developed to handle the subject of pronunciation variations in terms of acoustic modeling or extended lexicon and grammatical relations in terms of language modeling for non-native speech ASR. However, there are almost no methods that handle the difference between acoustic and linguistic features of non-native and native speech in a unified way, even if both features share a close relation and should be simultaneously taken into consideration.

Wang et al. (2016) proposed a novel phoneme set design method, based on the research results obtained with the previously proposed reduced phoneme set from the perspective of handling the acoustic and linguistic features of non-native speech in a unified way. The previously proposed reduced phoneme set was created with a phonetic decision tree (PDT) based top-down sequential splitting method (Wang et al. 2015a) that utilizes the phonological knowledge between mother and target languages and their phonetic features, delivering a better recognition performance for non-native speech. The recent approach considers acoustic and linguistic discrim-

inating performance in a unified way and optimizes the weighted total of both discriminating performances.

1.3 Examine the relation between proficiency of second language speakers and a reduced phoneme set customized for them

The proficiency of L2 speakers varies widely, as does the influence of the mother tongue on their pronunciation. As a result, the effect of the reduced phoneme set is different depending on the speakers' proficiency in L2. Wang et al. (2015c) examined the relation between proficiency of speakers and a reduced phoneme set customized for them.

The experimental results are then used as the basis of a novel speech recognition method using a lexicon in which the pronunciation of each lexical item is represented by multiple reduced phoneme sets (Wang et al. 2015b), and the implementation of a language model most suitable for that lexicon is described. Then multiple-pass decoding using a lexicon represented by multiple reduced phoneme sets is proposed for speech recognition of second language speakers with various prociencies. The relative error reduction obtained with the multiple reduced phoneme sets is 26.8% compared with the canonical one.

2 Future of Spoken Dialog Research

In the future of dialogue research, there are several challenges that are in need of attention. The first is how to collect enough and effective dialogical data for large scale statistical analysis, not only the adult speech and native speech, but also the children speech and non-native speech. The second is how to make better use of paralinguistic phenomena of human conversation characteristics, such as eye gazes, emotions and so on, to improve the performance of spoken dialogue systems.

I also work on collecting a multimodal corpus of human-to-human and human-to-robots conversations in English as second language by Japanese speakers in order to develop natural dialogue turn-taking model. Their voices were recorded together with the gaze-tracking data which is assumed to provide close cues for dialogue turns.

3 Suggestions for Discussion

- What's the core competence for basis research of dialogue systems in universities in comparison to that in industries?
- Standardization: How to collect enough and effective dialogical data for large scale statistical analysis? How to rapid development and standardization?
- Evaluation: How to evaluate the dialogue system (human-machine/robots)? Is there more standard one?
- Perspectiveness: How far can we go in deep learning for the research of dialogue systems?

References

Goh Kawai, and Keikichi Hirose. 1997. *A CALL system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruents*, EUROSPEECH, Rhodes, Greece.

Akinori Ito, Ryohei Tsutsui, Shozo Makino, and Motoyuki Suzuki, 2008. *Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system*, INTERSPEECH, pp. 2819–2822, Brisbane, Australia.

Xiaoyun Wang, Jinsong Zhang, Masafumi Nishida, and Seiichi Yamamoto, 2014. *Phoneme set design using English speech database by Japanese for dialogue-based English CALL systems*, LREC, pp. 3948–3951, Reykjavik, Iceland.

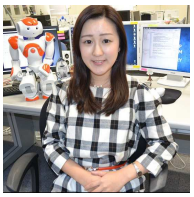
Xiaoyun Wang, Jinsong Zhang, Masafumi Nishida, and Seiichi Yamamoto, 2015. *Phoneme set design for speech recognition of English by Japanese*, IEICE Transactions on Information and Systems, 98(1): 148–156.

Xiaoyun Wang, and Seiichi Yamamoto, 2015. *Second Language Speech Recognition Using Multiple-Pass Decoding with Lexicon Represented by Multiple Reduced Phoneme Sets*, INTERSPEECH 2015, Dresden, Germany.

Xiaoyun Wang, and Seiichi Yamamoto, 2015. *Speech Recognition of English by Japanese using Lexicon Represented by Multiple Reduced Phoneme Sets*, IEICE Transactions on Information and Systems, 98(12): 2271–2279.

Xiaoyun Wang, Tsuneo Kato, and Seiichi Yamamoto, 2016. *Phoneme Set Design Considering Integrated Acoustic and Linguistic Features of Second Language Speech*, INTERSPEECH 2016, San Francisco, USA. (accepted)

Biographical Sketch



Xiaoyun Wang is a PhD candidate at the graduate school of Science and Engineering, Doshisha University, Kyoto, Japan, working under the supervision of Professor Seiichi Yamamoto. She is also a collaborative researcher in National Institute of Information and Communications Technology (NICT), Japan. Her research interests include speech recognition, spoken dialogue system, language acquisition, spoken language processing, and speech processing. She received a B.S. in Information Science from Yamanashieiwa University, Japan in 2012 and an M.S. from the graduate school of Science and Engineering, Doshisha University, Japan in 2014.

Ryan Lowe

McGill University
845 rue Sherbrooke
Montreal, Canada

ryan.lowe@cs.mcgill.ca
cs.mcgill.ca/~rlowe1

1 Research Interests

My primary research interests lie in the application of neural networks to building end-to-end dialogue systems. *End-to-end* refers to the fact that all the parameters of the model are jointly learned using a single objective function. This contrasts to the typical modular outlook on dialogue systems, where modules such as the natural language interpreter, dialogue state tracker, and natural language generator are learned separately. Note that we pursue this goal mostly in the context of chatbot-style systems, where the user has no specific goal.

One of the first to use such a paradigm in the context of neural networks was Ritter et al. (2011), who applied ideas from machine translation to predict the next utterance of a conversation on Twitter. Subsequent works have built upon this idea — now, more and more complicated neural network architectures are trained using maximum likelihood to predict the next utterance in a dialogue. The idea is that if a model assigns a high probability to the actual next utterance in a dialogue, it should be able to carry out a conversation reasonably well.

While I have been involved with creating some novel architectures for this problem (Lowe et al., 2015a; Serban et al., 2016), most of my work has focused on building *large datasets* that can be used to train such architectures, and finding ways to *evaluate* such architectures automatically when task completion rates are not available (as the user has no set 'task' to accomplish).

1.1 Large Dialogue Datasets

We first introduced the Ubuntu Dialogue Corpus (Lowe et al., 2015b), a large dataset of over 500,000 dialogues with an average of 7 turns. These dialogues consist of users on the Ubuntu IRC channel who are attempting to troubleshoot problems with their Ubuntu system. Other users respond and help the users to solve their problem based on their experience. While the original dataset is multi-party, we use a set of heuristics to disentangle the chat into dyadic (2-person) conversation. This dataset may be particularly useful for training models for technical customer service.

We also thoroughly surveyed the literature for existing datasets that could be used for building end-to-end dialogue systems (Serban et al., 2015), and found that there is a lack of very large datasets that can be used to train neural network models.

1.2 Dialogue Evaluation

Another direction of work is in trying to evaluate dialogue models automatically. While this has been well studied in the domain where task completion rates are given (such as the PARADISE framework (Walker et al., 1997)), it has been considerably less studied when such signals are not available. We proposed a recall-based method for evaluating these systems, called next utterance classification (NUC) (Lowe et al., 2015b), and analyzed how well humans performed on the task (Lowe et al., 2016). NUC evaluates how well a dialogue model can select the correct next utterance in a conversation, given a list of candidate utterances. This is a task that humans can perform quite well, while dialogue models still have some trouble, depending on the dataset.

We then investigated methods to evaluate the quality of responses generated by a model, given the context of a conversation. In particular, we examined the recent trend of using BLEU scores from machine translation to evaluate the quality of a response by comparing it to the ground-truth next utterance. We found that the BLEU score (as well as many other methods such as ROUGE, METEOR, and several word embedding based methods) correlate weakly or not at all with human judgement of the responses (Liu et al., 2016). Our current work is attempting to derive a system that can evaluate dialogue responses in a manner that correlates more significantly with human judgement of the response.

2 Future of Spoken Dialog Research

While there will always be a role for small dialogue systems in specialized domains, I believe that the next 5 to 10 years of dialogue research will bring about significant progress in building larger, end-to-end dialogue systems that can converse naturally about a variety of topics. Companies such as Google and Facebook are building large dialogue teams to tackle such problems using deep learning and reinforcement learning, and academia is sure to follow.

There are (at least) two large obstacles for such systems to be truly effective. The most immediate one is

the response diversity problem — current neural network models tend to produce fairly generic utterances compared to the diversity of human responses. This could perhaps be overcome using better architectures, larger datasets, or moving away from maximum likelihood (perhaps using a generative adversarial framework (Goodfellow et al., 2014)).

The second problem is how to incorporate natural language reasoning, or 'common sense', into these conversational agents. This is a much more complex problem, and will almost certainly not be fully solved in the next decade. Indeed, there is no theoretical ceiling to how complex we can make the reasoning of such agents. Progress in this area is crucial if we are to build agents that can converse naturally with humans.

3 Suggestions for Discussion

- Methods for automatic evaluation of chatbot responses (in a setting where the user has no specific task to accomplish)
- Desiderata for large conversational datasets, and candidates for such datasets
- Can we build intelligent conversational agents (e.g. generally intelligent chatbots) using only conversational data? Do we need visual grounding or other sources of multi-modality?

References

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, and Joelle Pineau. 2015a. Incorporating unstructured textual knowledge into neural dialogue systems. *NIPS Workshop on Machine Learning for Spoken Language Understanding*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015b. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *SIGDIAL*.

Ryan Lowe, Iulian V Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. On the evaluation of dialogue systems with next utterance classification. *SIGDIAL*.

Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*, pages 583–593. Association for Computational Linguistics.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. A hierarchical latent variable encoder-decoder model for generating dialogues. *arXiv preprint arXiv:1605.06069*.

Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. Paradise: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280. Association for Computational Linguistics.

Biographical Sketch

Ryan Lowe is a PhD student at McGill University, under the supervision of Joelle Pineau. Much of his work on dialogue is in collaboration with Iulian Serban, Yoshua Bengio, and Aaron Courville at the University of Montreal. He is also interested in reinforcement learning and causal models.

Ryan has previously worked at the Institute for Quantum Computing, the Max Planck Institute, and the National Research Council. His extra-curricular interests include rock climbing, reading fiction, writing, playing soccer, and hiking.

Chelsey Jurado

University of Texas at El Paso
Department of Computer Science
500 West University Ave.
El Paso, TX 79968

cnjurado@miners.utep.edu

1 Research Interests

Like many other introductory students to the computer science field, my expectations consisted of working on implementing algorithms and maintaining systems in an industry setting. Through various internship experiences, I have found myself yearning to do more. At the beginning of my junior year at the University of Texas at El Paso, I began working on research in speech and gaze features based on data collection. Later, I began to work with other students in a team on Embodied Conversational Agents (ECAs). So far, my work in research has shown me the field in which I will pursue my postgraduate work: intelligent systems in immersive environments. Though I do not know exactly what I will focus my future work in, my experiences have shown my interests to be in the study of adaptive conversations for agents.

1.1 Past Work

Under the guidance of Dr. Nigel Ward, I examined the correlation between prosodic features and gaze aversions in video chats (Ward et al., 2016). In dyadic conversations, it is known there are correlations between gaze aversions and speech patterns. In our work, we proposed a model which could predict when someone will be looking away from the interlocutor to improve video chat. The data included gaze behavior data in relations to the monitor, and the recorded conversations. We used leave-one-speaker-out training to create a preliminary model. At the best operation point, the model achieved 42% accuracy.

1.2 Present Work

Currently, my work in the lab focuses on ECAs and immersive environments. Users are able to communicate with the agent with the use of Microsoft Kinect's voice recognition and motion tracking engines. The agent is able to respond to both verbal and gesture responses. The system is, at present, dependent on the use of scripts consisting of what the agent is going to say and how to react verbally to an anticipated response. Our goal is to build rapport with the agent while improving the interaction with the agent. One project I am working on is to study how breathing can influence the interaction between the user and the agent.

1.3 Future Work

In my future work, I aim to improve our existing system by eliminating the system's dependency on scripts. The system would have some basic training from existing conversations, and with each user, will be able to learn and expand its knowledge base. The system would be given a simple goal, depending on the planned use of the agent, and the user will be able to have a conversation within that goal. This will allow for a more natural flow of conversation, and allow the experience to be more immersive and natural in its humanlike properties.

2 Future of Spoken Dialog Research

Dialog systems are becoming more conventional, for example, the rapidly increasing use of Siri, Cortana, and Google Now that have been improving in recent years. These agents are able to answer our questions, assist us in doing simple tasks, and some even have a sense of humor. These systems are limited to only accept a few commands, so as users, we learn to stick to these simple commands so that the conversations remain in a one-statement, one-response style. (Ward and DeVault, 2015) Within the next 10 years, dialog systems will still work to achieve a given goal, but they would be able to recognize a wider range of speech inputs, and will be able to respond accordingly. They will be more general in that they will be useful in multiple situations (e.g., tutoring, entertainment, providing information, etc.) given a goal. For this, as young researchers, we need to ask the following question, how can we efficiently combine learned and designed behaviors?

3 Suggestions for discussion

As a participant of the YRRSDS, I would be excited to meet and learn from the mentors and young researchers

that will be present. Four topics I would suggest to discuss at the event would be:

- Choosing your research topic,
- Possible ways to combine learned and designed behaviors,
- Challenges of open-world dialog, and
- Use of computer vision to manage floor control.

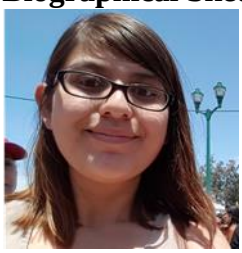
I believe that these discussion points will be both beneficial and interesting for everyone present.

References

Ward, N. G., Jurado, C. N., Garcia, R. A., & Ramos, F. A. (2016, March). On the possibility of predicting gaze aversion to improve video-chat efficiency. In Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications (pp. 267-270). ACM.

Ward, Nigel G., and David DeVault. "Ten challenges in highly-interactive dialog systems." AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction. 2015.

Biographical Sketch



Chelsey N. Jurado is currently working on receiving her B.S. in Computer Science from the University of Texas at El Paso. After earning her B.S., she plans on continuing her education and pursing a Ph.D. She is currently part of UTEP's Interactive Systems Group as an undergraduate research assistant. She joined the group in 2015 and worked under Dr. Nigel Ward. Currently she is part of the group's Advance Agent Engagement Team, working under Dr. David Novick. As a part of the Advance Agent Engagement Team she helps maintain and improve the embodied conversational agents created.

Adriana Carolina Camacho

The University of Texas at El Paso
Computer Science Department
500 West University Avenue
El Paso, Texas 79902
caro4764@gmail.com

1 Research Interests

My research interests lie in the **nonverbal conversational features** that may contribute to build **rapport** between humans and **Embodied Conversational Agents** (ECAs) in **multimodal** interaction. Some of these nonverbal features include task driven **gestures** and difference between **synthesized voice** versus **recorded voice.** More recently, I have concentrated my focus in the functions of **gaze** in dyad conversations and the importance of **breathing animations**.

1.1 Early Work

I started my research career at the Interactive Systems Group (ISG) in The University of Texas at El Paso (UTEP) During the first projects I participated, our research team analyzed the effects of synthesized speech and recorded across different virtual environments (Gris et al., 2014 The virtual environment either included 3D scenery appropriate for the game's nature or not included any scenery at all. In both conditions the ECA is shown standing in front of the user. We measured engagement by tracking gaze away from or toward the ECA and results indicated that a 3D environment and recorded speech had higher user-to-agent gaze time. In addition, we measured the times the agent was interrupted, which showed that the version with recorded speech had fewer interruptions than the versions that had the synthesized voice.

Given these results, our team's effort shifted to developing an engaging interaction that had a comprehensive story that surrounded an agent. To create more engaging agents we did a survey to list the features that make a "perfect" agent. Items of interest included the potential need for an agent's persona and anthropomorphism (agents with human-like characteristics in terms of behavior and appearance).

In a following study (Camacho et al., 2014) we reviewed the quality in terms of graphical representation, believability, and overall naturalness of existing, commercial, or available ECAs by identifying and evaluating their relevant features. We identified two main characteristics: visual and functional and explored the changes of these qualities as a function of time across several ECA versions when appropriate (i.e. the same agent appeared in several representations, with additional features or improvements over time). Respectively, we identified five key visual features and eight key functional features. However, we acknowledged that a better rubric could be based on a study of user's perceptions that explores the relative importance of the features and whether a feature outweighs another (e.g. a realistically human-like but unusable agent).

After identifying these features, we developed more advanced agents in terms of dialog and visual appeal to engage our users and potentially lead us to results that might indicate higher rapport building while having fully automated agents. The result was "Survival on Jungle Island" (Novick et al., 2016) game, where users carry conversations and a series of activities led by our ECA. We simulated a survival scenario so the user could cooperate with the ECA and create rapport-building opportunities. In addition, we wanted to take advantage of the non-verbal behaviors in a more immersive environment and perform task driven gestures. Since the experiment was devised to measure the effectiveness of gesture-enabled interactions on rapport, two versions were developed to compare participants' experiences between what we denominated the gesture and non-gesture enabled agent versions. To facilitate in gesture analysis and annotation, our system included a module where we capture a collection of poses to be recognized as gestures in the human-ECA interaction and automatically annotate gestures performed by the user using a Microsoft Kinect (Gris et al., 2015). Subjects were previously tested to find their personality-type, enabling us an initial body of recordings for analyzing gestures differences between extraverts and introverts when interacting with an ECA. The goal was to analyze the paralinguistic behavior, and measure the correlation of these behaviors with rapport building. In addition, these behaviors are not limited to facial expressions, hand gestures, or task-related gestures, but rather a set of normal, unconscious gestures and poses over a relatively long period of interaction (40 minutes to an hour). However, these experiments were cumbersome to setup, leading to our current advancements.

1.2 Current and future focus

My current research focus centers on developing automated interactions between humans and virtual agents. In addition, I work on building the infrastructure that

makes these interactions possible, by developing features that control several parts of the agent's behavior and facilitate ECA development. We are developing a systematic way of creating these interaction by creating a scripting system and a simple dialog manager, both of which are now modules of the UTEP AGENT Framework (Gris et al., 2015). We are also working on a tool that will enable users create their own human-ECA interactions as if they were writing a movie/play script.
Lately, I have participated in a shared project with the Communications Department where the aim is to engage humans with cultural aspects (through the use of traditional cuisine) using multiple ECAs.

I am now exploring gaze as a turn-taking modulator. One of the problems in the "Survival on Jungle Island" was that users tended to interrupt the agent by failing to notice the almost non-existing turn-taking mechanisms. Adding proper gaze aversion to our agents helps the user know when to speak, leading to less interruptions.

I am also exploring other non-verbal features that play key roles in dialog naturalness and believability such as breathing. Our intention is to find out whether this will play a positive/rapport building role in the interaction, or will it damage it. Does likeability mean that breathing agents are better?

2 Future of Spoken Dialog Research

I believe that this generation of researchers will be able to take steps towards a dialog system that adapts to its users to better accomplish their goal, and more specifically doing it in multimodal systems. One possible track of doing so is by looking at the personality types and analyzing the differences of users in different domains. Cultural background could also play a big role in adaptability and therefore also presents a potential problem to be analyzed and solved.

3 Suggestions for discussion

- Potential of dialog systems in Augmented Reality, Virtual Reality, and Mixed Reality.
- Influence of different personality type adaptations in dialog systems.
- The use of prosody in dialog systems to convey different messages using the same words and the importance of it.

References

Camacho, A., Alex Rayon, Ivan Gris, and David Novick. "An Exploratory Analysis of ECA Characteristics." In *International Conference on Intelligent Virtual Agents*, pp. 95-98. Springer International Publishing, 2014.

Gris, Ivan, David Novick, Adriana Camacho, Diego A. Rivera, Mario Gutierrez, and Alex Rayon. "Recorded speech, virtual environments, and the effectiveness of embodied conversational agents." In *International Conference on Intelligent Virtual Agents*, pp. 182-185. Springer International Publishing, 2014.

Gris, Ivan, Adriana Camacho, and David Novick. "Full-body gesture recognition for embodied conversational agents: The UTEP AGENT gesture tool." In *Gesture and Speech in Interaction, Nantes, France*, 2015.

Novick, David, Iván Gris, Diego A. Rivera, Adriana Camacho, Alex Rayon, and Mario Gutierrez. "The UTEP AGENT System." In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 383-384. ACM, 2015.

Novick, David, Adriana Camacho, Ivan Gris, and Laura M. Rodriguez. "Tracking Conversational Gestures of Extraverts and Introverts in Multimodal Interaction." In *RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (May 28 th, 2016)*, p. 24.

Biographical Sketch

Adriana Camacho obtained her bachelor's degree in Computer Science in 2014 from UTEP, where she currently is pursuing her PhD. Her research focuses on the development of rapport-building embodied conversational agents in ISG. She is a recipient of the National Science Foundation Graduate Fellowship program. She has co-authored and authored in the International Conference on Intelligent Virtual Agents, Gesture and Speech in Interaction Conference, International Conference on Multimodal Interaction (ICMI), and the RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop. She is part of the team that won best demonstration at the ICMI 2015.
Adriana was invited as a PhD student panelist for the Computing Alliance of Hispanic-Serving. She has also managed outreach demonstrations to high school, middle school, and elementary school students and is part of the Inmerssion start-up company that works with virtual reality and immersive environments.

Tiancheng Zhao

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA, USA, 15213

tianchez@cs.cmu.edu
www.cs.cmu.edu/~tianchez

1 Research Interests

My research interest in general lies in sequential decision making for spoken dialog system, such as turn-taking and dialog management using reinforcement learning. Specifically I am interested in pursing scalable, flexible and end-to-end trainable computational model that jointly optimize the decision making process of a conversational agent that can carry out meaningful dialogs with human users in multiple-domains. The following sections briefly introduce the research projects that I worked on in the past and the ones that are in my current and near future research focus.

1.1 Active Turn-taking with System Barge-in

One essential functionality of mixed initiative dialog is allowing users to interrupt systems utterances. Naive system interruption to users may be considered as inappropriate, recent studies, however, have shown that clever system barge-in can actually improve the task success rate (Ghigi et al., 2014). Our work (Zhao et al., 2015) investigates the effectiveness of active system barge-in in a more principled way by incorporating reinforcement learning and the theory of optimal stopping. The study first views a user utterance as a stream of partial hypotheses from the speech recognizer and constructs a cost model based on the number of correct and incorrect slots in each partial. Two possible actions: continuing listening and active barging in, are available to the the machine. Then the optimal action at each partial hypothesis is calculated according to the theory of optimal stopping and the authors apply a Support Vector Machine (SVM) to learn the optimal policy.

Results show that the trained classifier is able to correctly estimated the oracle actions with a F-measure of 85.9%. Moreover, the simulation results demonstrate that active system barge in can improve the average reward by 27.7% in the best case and is able to estimate user intention approximately in the middle of utterances. This suggests that machine should have its own motivations to actively take actions that are beneficial for the task successfulness.

1.2 DialPort: Connect the SDS community

DialPort[1] is one of the long-term project that I am focusing on. The basic idea behind DialPort is to construct a shared platform for the academic SDS community to collect real-user data and experiment with new systems. DialPort is formulated to be a multi-agent dialog system that each agent is a independent dialog system from external research labs. For the actual users, however, they observe a unified interface as if talking to a single virtual agent. Currently, DialPort has successfully connect to three external systems, including yelp API, a chat bot and the one form the Cambridge University. In the future, we will continue to conduct research in better ways of connecting various dialog systems and corporate with other research labs over the world to make DialPort more powerful. For more information, please refer (Zhao et al., 2016) and we sincerely welcome anyone interested in to join DialPort.

1.3 End-to-End Learning for Dialog State Tracking and Management

The second project in my current focus is developing end-to-end trainable task-oriented dialog system (Zhao and Eskenazi, 2016). Inspired by recent success in deep reinforcement learning (DRL), the proposed approach views the slot-filling as a part of the actions for a reinforcement learning agent and therefore jointly optimizing the policy for both dialog state tracking and dialog policy. Further, we also use the expressive power of recurrent neural networks (RNN) to automatically learn dialog state representation along with policy learning. This results in an end-to-end trainable model given the dialog success and slot-filling labels.

The experiments showed that the proposed algorithm is able to master in playing a conversational game with simulated user and our model analysis indicated compelling evidences that the embedding layer of the RNN is learning important state feature of a dialog. However, we also observed that the proposed model suffers from poor sample complexity, which will be investigated in the future work.

[1] https://skylar.speech.cs.cmu.edu

2 Future of Spoken Dialog Research

I believe that we are in the golden age of spoken dialog research since there are significant more attention to our field in recent years from both industry and academia. One of the reasons I believe is that the technologies of automatic speech recognition (ASR) and natural language processing (NLP) have evolved well enough so that it becomes plausible to build software that can carry out more complex reasoning and interaction with human users.

As one of the young researchers, I think we should pursue both practical and theoretical research. From practical perspective, constructing scalable development framework can allow both companies and research labs to better prototype various dialog systems, such as personal assistant, virtual tutor and many others. Meanwhile, it is important for researchers to pursue theoretical foundation that connects the bridge between traditional discourse theory with practical dialog development framework. Such theory is crucial to produce principled study in the field in the long term.

3 Suggestions for Discussion

I would like to suggest the following topics for discussion.

- **Evaluation of dialog systems:** Dialog system evaluation is known to be challenging. In other AI areas, such as translation or computer vision, that simple accuracy or BLUE score, though not perfect, provides a common metric for system comparison. Exploring novel methods to compare similar dialog system is a important and interesting topic for discussion.

- **Categorization of Dialog Systems:** Numerous types of systems have been proposed to conduct conversation with human users. Each of them uses different approaches and serve for various purposes. However, there is a lack of accepted categorization that classifies existing dialog systems into well-defined types with respect to their approaches or functionality. This would not only be helpful in identifying the unique and common challenges in each category, but also valuable for young researchers to obtain the "big picture" of spoken dialog research.

- **End-to-End Dialog Systems:** Recently, there have been increasing interest in developing end-to-end trainable dialog systems (Iulian et al., 2015; Wen et al., 2016; Zhao and Eskenazi, 2016; Williams and Zweig, 2016). An interesting discussion could be about the pros and cons of various approaches and the challenges that observed in these studies.

References

Fabrizio Ghigi, Maxine Eskenazi, M Ines Torres, and Sungjin Lee. 2014. Incremental dialog processing in a task-oriented dialog. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Tiancheng Zhao, Alan W Black, and Maxine Eskenazi. 2015. An Incremental Turn-Taking Model with Active System Barge-in for Spoken Dialog Systems. In *Proceedings of the SIGDIAL 2015 Conference*.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. In *Proceedings of the SIGDIAL 2016 Conference*.

Tiancheng Zhao, Kyusong Lee and Maxine Eskenazi. 2016. DialPort: Connecting the Spoken Dialog Research Community to Real User Data. arXiv:1606.02562v1.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. arXiv:1604.04562.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. arXiv:1507.04808.

JD WilliamsG Zweig. 2015. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. arXiv:1606.01269v1

Biographical Sketch

Tiancheng Zhao is a first year PhD student at Language Technologies Institute, in School of Computer Science Carnegie Mellon University, working under the supervision of Prof. Maxine Eskenazi and Prof. Alan W Black. His research interest lies in spoken language processing and sequential decision making for dialog systems. Currently he focuses on combining dialog expert knowledge and deep reinforcement learning, in order to efficiently develop end-to-end multi-domain conversation models that can conduct both goal driven and non-goal driven dialogs with human users. Previously, Tiancheng received his M.S in Language Technologies from Carnegie Mellon University in 2016. Prior to that, he received the B.S. in Electrical Engineering from University of California, Los Angeles in 2014, advised by Professor Abeer Alwan.

Cassidy Rae Henry

Student UCLA // Visiting Research Assistant ARL
384 Walnut Ave
Long Beach, CA 90802

cassidy.henry@ucla.edu
www.shapkaa.com

1 Research Interests

My developing research interests are primarily in **linguistics**. I especially enjoy **computational linguistics**, and in addition to using computational analysis within linguistics, I am very interested in **artificial intelligence, machine learning, and neural networking**. My favorite subfields of linguistics are **morphology, phonology, phonetics,** and social/behavioral **affect in language,** especially in the context of **human-robot interaction**.

1.1 Present work

Currently, I am doing work on a robot-based dialog system at ARL. I am researching to investigate how various perceptions of a robot teammate may affect one's language or paralinguistic speech towards a robot. This is my first dialog system project, but I am very excited to continue research in this area, as it ties in very closely with my strong research interests in artificial intelligence and computational linguistics. I have also participated in several data collection experiments for USC ICT's New Dimensions in Testimony project, and attend their weekly Natural Language Group meetings, which is furthering my interest in dialog systems research. It is inspiring to be able to work with top people in the field on projects such as these.

Outside of dialog research, I am currently working on a illustratory phonology project of a Turkic language spoken in Eastern Europe/Russia with another student in my department under the supervision of a faculty member. This project is an extension of a final paper I completed for a phonetics class. We are hoping to submit it for publication in the JIPA.

In terms of the future, I also have many personal research projects that I hope to have more time to develop, such as large-scale analysis of Internet linguistic phenomenon, amongst other linguistic topics. My home department also highly encourages research projects within their classes, so I will also be doing lots of linguistics research there too.

1.2 Past work

In the past, I have done corpus-based linguistics projects, such as analysis of loanwords in the Russian language in terms of when they entered the language and how frequently they are used. In that project, I found that words relating to technology and other modern phenomena tended to enter in the post-Soviet era, and words otherwise experienced a vast increase in usage relative to other words at the time in the post-Soviet era.

2 Future of Spoken Dialog Research

Dialog research in the next few years will most likely be focusing on automation. Instead of humans building corpuses or question-answer systems for dialog systems, dialog systems in the future will be based off of intelligent natural language systems that are trained to handle specific information, similar to how IBM Watson has shown it can be used, using Wikipedia as its "knowledge base." In the time between then and now, young researchers will be the ones working on the solutions to the problems standing in the way between now and then. I believe with the huge strides of progress being made within computational linguistics currently, that dialog systems research will as a result have great progress.

The current problems that need to be addressed are wide in scope. One of them is of particular interest to me – how affect and paralinguistic elements can be integrated into dialog systems, particularly in the forms of detection and understanding. Another big consideration is including paralinguistic information such as gestures, which will allow for future accessibility (laying a pathway for interpreting gesture could allow dialog systems that use sign languages, leading to greater accessibility of the dialog system. These are pertinent issues that should be considered while other big issues are being evaluated and rectified in dialog systems research.

Other issues include addressing linguistic variation, allowing for wider sets of behavior, state tracking, and more. David DeVault of ICT and Nigel Ward

describes these issues in great detail in *Ten Challenges in Highly-Interactive Dialog Systems.*

3 Suggestions for discussion

In terms of potential discussions at YRRSDS 2016, I would be interested in a variety of topics. One such topic I think of often is how to incorporate facial expressions and gesture-based communication into dialog systems. This is not only important because of our very gesture-rich and expressive movements in communication, but also for accessibility – people who are hard of hearing should be able to use sign language as an input. ASL speakers don't just sign blank-faced, however. The prosodic elements of speech for ASL are strong – from their facial expressions, to rate of movement and even flow in movement, there's lots going on. While things like this could be identified in vocal speech ASR through analysis of soundwaves (and is currently underdeveloped in research), it is much harder to recognize prosodic elements of sign-based speech for an AI system.

This also broaches the broader topic of accessible dialog systems. It is important for researchers to address the issues of accessibility now while dialog systems are in development, before they are implemented on a wide-scale basis and then cannot be used by an entire population group due to some issue of accessibility. Dialog systems also have great potential for expanding access of information or making certain tasks easier for people with certain kinds of disabilities, so it is important to be accessible in that regard as well.

Another area of interest I find worth discussing is automation of grounding techniques in dialog systems. Grounding is the strongest way that one can show that they are both following a conversation and understanding what the other conversational party is talking about. Therefore, a dialog system that automatically uses grounding techniques could be viewed as more efficient or trustworthy based upon how good it is at doing so. In certain context, the user's perception of the dialog system can matter a lot in how successful they are at completing a task. This falls under the umbrella of natural dialog systems.

With the explosion of neural networking and machine learning within computational linguistics, I think it would be fun to consider expansive techniques and current research problems for natural dialog systems. I believe that while efficient for simple tasks, dialog systems with limited vocabulary are not engaging. Natural dialog systems allow the user to actually *engage* the conversational agent instead of just being required to adapt to a question-answer system. Many exciting research problems currently exist within the field of natural dialog systems and this would be great conversation at YRRSDS, as young researchers are the ones who will be working on these issues.

References

Dialog. (n.d.). Retrieved July 15, 2016, from http://www.ibm.com/watson/developercloud/dialog.html

Nigel G. Ward and David DeVault. 2015. *Ten Challenges in Highly-Interactive Dialog Systems.* Institute for Creative Technologies, University of Southern California.

Biographical Sketch

Cassidy Henry is a 21-year-old, current senior undergraduate student at the University of California, Los Angeles (UCLA) studying Linguistics with a specialization in Computer Science, and a minor in Russian language. Prior to that, she was enrolled at College of the Canyons to prepare for transfer to UCLA. In the past, she has also studied Russian at Kazan National Technological University in Kazan, Tatarstan, Russia on the U.S. Department of State's National Security Language Initiative for Youth (NSLI-Y) program. She plans to pursue a PhD in Linguistics following her undergraduate career, while maintaining the focus on computational linguistics. Cassidy aspires to work as an academic researcher upon completion of her PhD, as she finds research invigorating and exciting.

Cassidy is also currently a visiting research assistant at the U.S. Army Research Lab, West (ARL West) campus, where she works on a bot language project under the supervision of a team in the Human Research and Engineering and Computational and Information Sciences Directorates.

Bing Liu

Carnegie Mellon University
NASA Research Park, Bldg 23
Moffett Field, CA 94035

liubing@cmu.edu

1 Research Interests

My main area of research lies in **Natural Language Understanding (NLU)** and its applications in **Spoken Dialogue Systems (SDS)**. I am particularly interested in **knowledge representation** and integrating knowledge to dialogue systems. My initial research focus is on distributed representation of text and its applications in spoken language understanding. My recent efforts are more towards learning distributed representation of knowledge and context. I am keen on exploring knowledge-enabled end-to-end trainable dialogue systems that can perform tasks with specific goals.

1.1 Text Representation and SLU

Spoken language understanding (SLU) system interprets the semantic meanings conveyed by speech signals. Major components in SLU systems include identifying speaker's intent and extracting semantic constituents from the natural language query, two tasks that are often referred to as intent detection and slot filling. Intent detection can be treated as a semantic utterance classification problem, and a number of standard classifiers can be applied. Slot filling can be treated as a sequence labeling problem. These two tasks are usually processed separately by different models. Prior work on intent classification and slot filling includes using uni-bi-trigrams, together with head word, hypernym, and other hand-coded rules in classifiers like SVM.

Motivated by the success of distributed representations of words in many NLP applications (Collobert et al., 2011), we designed an RNN-based SLU model with scheduled training method (Liu and Lane, 2015) that can effectively capture the sequential structural patterns of the text. In a following-up work (Liu and Lane, 2016a), we developed an RNN-based model that can be used to jointly perform intent detection and slot filling in SLU. Such joint models simplify SLU systems, as only one model needs to be trained and deployed. In this work, we also explored the effectiveness of applying neural attention mechanism in the RNN model, and evaluated the attention based RNN model in sequence-to-sequence learning settings.

In many real world applications using speech interface, real time responses from the agent are desired. In speech recognition, instead of receiving the transcribed text at the end of the speech, users typically prefer to see the ongoing transcription while speaking. In SLU, with real time intent identification and semantic constituents extraction, the downstream systems will be able to perform corresponding search or query while the user dictates. In a recent work (Liu and Lane, 2016b), we designed an RNN-based online joint SLU model that keeps tracking the intent variations as word in the transcribed utterance arrives and uses it as contextual features in the joint model. In addition, we modeled the interaction between the SLU and the language modeling tasks and showed that such joint modeling led to better ASR and SLU performance. As a next step, we want to see how such joint model can be further extended for belief tracking in dialogue systems when considering the dialogue history beyond a single utterance.

1.2 Knowledge Representation

Introducing knowledge to dialogue and question answering systems is an important step towards building intelligent conversational agents. Knowledge comes in various forms, and how these knowledge can be represented and integrated to the intelligent system in an end-to-end trainable manner is an active research problem. Large scale knowledge bases are good sources for structured knowledge. Such knowledge is entity centric and is typically represented by properties and relations between entities. Embedding methods in representing the entity and relations has been explored in literature (Socher et al., 2013). Key challenges in question answering with knowledge base include the very large search space in knowledge base and query compositionality. Knowledge in unstructured form can come from web documents (e.g. text and images) that are returned by search engine given a query. Knowledge from web document is likely to be more complete and updated comparing to that from knowledge base. The main challenges are the accurate understanding of the query and knowledge extraction from data in multi-model form.

My current research interest lies in extracting and integrating task-specific knowledge to the dialogue manager. To control the search space, we want the knowledge to be dynamically retrieved from knowledge base or web doc-

uments based on the task and detected user intent.

1.3 Task-oriented Dialogue Systems

Sequence to sequence models have shown promising improvement in building end-to-end trainable, open domain dialogue systems (Vinyals and Le, 2015). Such models typically require a large amount of data to train, which is often not available for task-oriented applications. To mitigate this problem, popular approaches in closed domain dialogue system design cast the task as a partially observable Markov Decision Process (POMDP) and use reinforcement learning for online policy optimization by interacting with users (Gašić et al., 2013). Recently, a neural network based end-to-end trainable task-oriented dialogue system (Wen et al., 2016) is proposed with a novel crowdsourced data collection framework and showed promising performance. In my research, I would like to whether the problem of insufficient training data in task-specific end-to-end trainable dialogue systems can be mitigated by the effective usage of the introduced knowledge.

2 Future of Spoken Dialog Research

In the long run, I think a truly intelligent conversational agent will not only be able to handle comprehensive queries, but also to understand human emotions and express personality. The agent should be able to understand complex questions, provide solutions with updated real world knowledge, and respond in a manner that is adjusted based on the user's characteristics.

In the near future, I believe we will see the growing importance of data-driven approaches in designing dialogue systems, in both open and task-oriented domains. Current spoken dialogue systems contains large amount of handcrafting, and this substantially limits the systems' domain scalability. Data driven approach with customized domain knowledge can be a promising solution. Moreover, I see good potential of applying transfer learning in spoken dialogue systems. It will be interesting to see how a well performing task-specific dialogue model can be efficiently adapted to a new domain that with limited training data.

3 Suggestions for Discussion

I would like to suggest the following topics for discussion.

- Integrating knowledge or external memory to dialogue systems.
- Transfer learning in task-specific dialogue systems.
- Evaluation: what can be good metrics in evaluating the responses generated by conversational agent?
- What are the main challenges in deploying real world spoken dialogue systems?

References

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

M Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013. On-line policy optimisation of bayesian spoken dialogue systems via human interaction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8367–8371. IEEE.

Bing Liu and Ian Lane. 2015. Recurrent neural network structured output prediction for spoken language understanding. In *Proc. NIPS Workshop on Machine Learning for Spoken Language Understanding and Interactions*.

Bing Liu and Ian Lane. 2016a. Attention-based recurrent neural network models for joint intent detection and slot filling. In *INTERSPEECH*.

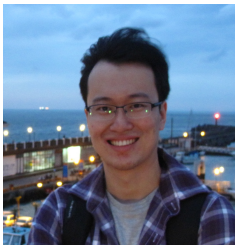
Bing Liu and Ian Lane. 2016b. Joint online spoken language understanding and language modeling with recurrent neural networks. In *SIGDIAL*.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Biographical Sketch



Bing Liu is a PhD student in the Department of Electrical and Computer Engineering at Carnegie Mellon University, working under the supervision of Professor Ian Lane. His research interests include using machine learning for natural language processing, spoken language understanding, and dialogue systems. Bing graduated with a Bachelor degree in Electrical and Electronic Engineering from Nanyang Technological University (Singapore).

Maike Paetzel

Uppsala University
Box 337
SE-751 05 Uppsala
Sweden

maike.paetzel@it.uu.se
http://www.maike-paetzel.de/

1 Research Interests

Conversational agents in social contexts need more than optimized task performance skills: In order to keep the human dialogue partner engaged in the conversation, agents need to find a good **balance between task oriented and social behavior**. Different dialogue policies and interaction strategies can highly influence the user's perception of a dialogue system. During my Master thesis project, I was mainly looking into **different dialogue strategies in incremental vs. non-incremental dialogue systems** and how they can influence both task performance and user satisfaction with the system.

My PhD in social robotics shifted the focus more towards interaction strategies for embodied conversational agents. The goal is to investigate how **different embodiments, appearances and interaction strategies can influence the perception of the agent** in a human interaction partner. Mainly, I want to investigate what in dialogue policies can cause a strange or uncanny feeling towards the conversational agent, which ultimately undermines the goal of the conversational agent to act as a social companion.

1.1 Incremental Dialogue Systems and Optimized Policies for Rapid Conversations

As part of the NSF funded project "Incremental Speech Processing for Rapid Dialogue" we developed a conversational agent capable of playing a rapid object matching game, which could demonstrate the values of a fully incremental system architecture and a dialogue policy optimized towards the incremental nature of the system.

The Domain To enforce a conversation with fast turn-taking behavior between dialogue partners, we developed the Rapid Dialogue Game Image (RDG-Image) (Paetzel et al., 2014). In this game, two players need to cooperate in order to score as high as possible in a given time limit. Both see the same set of eight images on their screen in shuffled order. One of the players acts as a director, who sees the target image highlighted on the screen. The director now describes the image verbally to the matcher, who needs to select the same image on the matcher's own screen. Once the matcher is confident about the selection, the director can request the next target image.

Corpora Two corpora were collected including two human partners playing the game. The RDG-Image Lab corpus was recorded on-site at the Institute for Creative Technologies, University of Southern California, and includes 32 pairs of gameplay (Paetzel et al., 2014). As this data collection was cost and time intensive, Manuvinakurike and DeVault (Manuvinakurike and DeVault, 2015) developed a web framework to crowdsource the data collection. On Amazon Mechanical Turk, 98 human-human game plays could be recorded for the RDG-Image Web corpus.

The Eve Agent During my Master thesis project, we developed an automated agent called **Eve** which is able to autonomously play the role of the matcher in the RDG-Image game. We designed a dialogue system from scratch which allows to easily switch between incremental and non-incremental processing in different modules of the system. By testing a fully incremental version of the system against a semi-incremental and non-incremental version, we could highlight the value full incremental processing adds both to the game performance and user satisfaction (Paetzel et al., 2015)(Manuvinakurike et al., 2015b). In addition, we could show that a dialogue policy optimization which is fitted towards the incremental nature of the architecture is crucial to use the full possibilities of the dialogue system (Paetzel, 2015).

1.2 Interaction Strategies for Embodied Conversational Agents

In my PhD, the focus of my research shifted towards interaction with embodied agents. One main goal of my thesis is to investigate how different interaction and dialogue strategies can cause or prevent an uncanny or eerie feeling in humans. During my first year, I conducted a series of pilot studies to understand the perception of the robot platform Furhat in general and the perception of gender in particular. Furhat is a back-projected robot head, which allows to model facial features and expressions in an easy and cost-effective manner. In a first pilot study, we could show that the robot platform is indeed suitable for research on uncanny feelings and multimodal interactions (Paetzel et al., 2016). Especially, visual incongruent gender cues raised an uncanny feeling towards

the robot in adults. In a second study, we used incongruent multimodal gender cues in a short interaction with children. Here, for various reasons, we did not find a link between incongruence and uncanniness. However, we observed that children rely more on auditory than on visual cues when judging the perception of a robot, which supports future plans of investigating different dialogue strategies and their effect towards the feeling of uncanniness.

2 Future of Spoken Dialog Research

Up to date, most end users still prefer commanding their computer devices by keyboard, mouse or touch input, even though the usability of state-of-the-art dialogue systems, for example in mobile applications, have improved during the last years. However, since users have a different option of interaction which they are familiar with for years, the scepticism towards language interaction with computers is still high. With the raise of robots in domains like household and health care, the demand for natural interaction with computer systems will likely increase in the next five to ten years, as robots don't offer another option of interaction than spoken language, mimic and gesture.

This development can be boon and bane at the same time: When dialogue research becomes much more popular, a more natural interaction with machines might be achieved quicker. On the contrary, interaction with robots adds huge challenges to the field of dialogue systems:

- Instead of dialogue narrowed down to a specific use-case or domain, open-domain dialogue abilities will be required.
- While many current systems use massive computation powers on server farms to process spoken input, people might have privacy concerns if their robot is sending all microphone input from a home environment to external processing. In addition, good internet connections are still not available everywhere. For this reasons, the trend might move back towards better offline algorithms with limited resources.
- Personalized individual robots will require personalized dialogue behaviour, with the ability to memorize preferences of multiple people over the period of years.
- As natural interactions prohibits the use of headsets, dialogue systems will need to become much more robust towards background noise.

3 Suggestions for Discussion

As my main focus is in multimodal interaction with embodied agents, I would be interested in discussing the following questions:

- The influence of embodiment: How does the type of embodiment (2D vs 3D) and the appearance influence the language and dialogue strategies humans use in their conversation with the dialogue system?
- Evaluation: How can we find standardized metrics to evaluate conversational agents across domains and embodiments?
- How can incremental processing be scaled up to processing and generating multimodal interactions?

References

Ramesh Manuvinakurike and David DeVault. 2015. *Pair Me Up: A Web Framework for Crowd-Sourced Spoken Dialogue Collection*. Natural Language Dialog Systems and Intelligent Assistants.

Ramesh Manuvinakurike, Maike Paetzel and David DeVault. 2015. *Reducing the Cost of Dialogue System Training and Evaluation with Online, Crowd-Sourced Dialogue Data Collection*. The 19th workshop on the semantics and pragmatics of dialogue (SemDial).

Maike Paetzel, David Nicolas Racca and David DeVault. 2014. *A Multimodal Corpus of Rapid Dialogue Games*. Language Resources and Evaluation Conference (LREC).

Maike Paetzel, Ramesh Manuvinakurike and David DeVault. 2015. *"So, which one is it?" The effect of alternative incremental architectures in a high-performance game-playing agent*. Special Interest Group on Discourse and Dialogue (SIGDIAL).

Maike Paetzel. 2015. *Exploring the effect of incremental speech processing on dialogue policy performance in a game-playing agent*. Master thesis.

Maike Paetzel, Christopher Peters, Ingela Nystrm and Ginevra Castellano. 2016. *Preliminary results from using a back-projected robot head in uncanny valley research*. Extended abstract at the RO-MAN Interactive Session.

Biographical Sketch

Maike Paetzel just finished her first year as a PhD student at Uppsala University, Sweden, under the supervision of Ginevra Castellano. In her Bachelor and Master studies at the University of Hamburg, Germany, she focused both on robotics and embodied agents as well as dialogue systems. In her PhD in social robotics, she is now working on bringing the two fields of interest together.

Alexandros Papangelis

Toshiba
Cambridge Research Lab
Speech Technology Group

alexandros.papangelis@crl.toshiba.co.uk

1 Research Interests

Within the broad area of Statistical Spoken Dialogue Systems, my research interests lie on **online adaptive dialogue management**, on **socially skilled agents**, and on **multi-domain dialogue**. Regarding dialogue management (DM), during my PhD I have specifically focused on applications of Robot Dialogue Systems (RDS) for Assistive Living Environments (ALE), where the users are by definition not competent with technology. I then moved on to socially skilled virtual humans, focusing on building and maintaining long-term rapport with users, for education (peer tutoring) and other applications. My current focus is on multi-domain statistical dialogue management, using methods that can scale well.

1.1 Online Dialogue Management

Reinforcement Learning (RL) has been used for many years in dialogue systems research and has many benefits, including the fact that there is a plethora of mature and proven algorithms, many of which are scalable. During my PhD, we performed extensive evaluations of standard and state of the art (at the time) RL algorithms for dialogue management. We then applied these methods on adaptive RDS, in ALE scenarios. Having a physical robot adds a whole new dimension, impacting many aspects of the interaction, for example proxemics, gaze, gestures, etc. ALE pose even more challenges, including the fact that users in such environments have difficulties interacting with standard technology interfaces, either due to physical (injuries, age, chronic conditions etc.) or mental disabilities and the vast amounts of available data, such as real-time streams from sensors that need to be fused, interpreted and stored in an efficient manner.

To address some of these challenges, we proposed two batch learning algorithms, able to learn how to combine basic system actions into complex ones and achieve mandatory system goals (such as rehabilitate user) while achieving as many optional goals as possible (such as entertain user) (Papangelis et al., 2012). Our next step was to create a system able to help users suffering from psychological disorders such as Post-Traumatic Stress Disorder (PTSD) (Papangelis et al., 2013; Papangelis et al., 2014a). We then used a similar system to collect multimodal data from human users (Tsiakas et al., 2015).

Other work of mine focused on dialogue management using RL to learn good two-party multi-issue negotiation dialogue policies (Papangelis and Georgila, 2015). In this work, we used a simulated user for training an RL agent and human users for evaluation. Our simulated user extends the agenda-based paradigm to fit the negotiation scenario. In this scenario, each negotiator has goals and preferences as well as strong and weak arguments for each issue and can make a variety of moves, for example propose an offer or trade-off, provide an argument, accept or reject an offer, etc.

1.2 Socially Skilled Agents

Recently, I have worked on intelligent virtual agents that employ sociocultural models in order to build or maintain rapport between themselves and human users and thus more efficiently perform tasks. This work is in line with a broader research area that I am interested in: socially skilled agents, i.e. agents that behave in a socially acceptable manner. There are many challenges involved in such efforts, including recognising the social intentions behind the user's behaviour, responding in a manner that achieves the system's social and task related goals and properly realizing this response. To accomplish this, we did an extensive analysis of sociopsychological findings about how people build, maintain and break rapport and proposed a theoretical model of managing rapport between humans and virtual agents (Zhao et al., 2014). Based on this model, we proposed a computational architecture (Papangelis et al., 2014b).

To meet some of the challenges, I worked on a *social reasoner* that, similar to a DM, deliberates about the social dimension of the interaction and decides what is the best social strategy to follow in the next dialogue turn, and a *social state tracker*, a module that can map features extracted from the input (audio / video) into the social functions that lead to rapport (mutual attentiveness, face management, coordination) and keep track of the dynamics of these social functions, following our proposed architecture (Papangelis et al., 2014b).

My work with virtual humans focused on two projects, a virtual peer tutor and a mobile personal assistant. While the social goals in both systems are similar (build and maintain rapport over the long-term), the task is quite different: the first system is a teenager who can teach and be

taught and the second is a personal assistant that connects the user to Yahoo! services.

1.3 Multi-domain Dialogue Management

I am currently working on methods that allow information-seeking dialogue systems to operate across multiple domains. More specifically, I am investigating how we can apply transfer learning methods to design a multi-domain dialogue manager.

2 Future of Spoken Dialogue Research

There is an increasing number of interactive systems (embodied or not, virtual or physical) that take sociocultural norms into account. There is also an increasing number of works that use Neural Networks and Deep Learning to handle the complexities of human conversation. In the near future, I expect these to intersect and create socially skilled agents that can handle lifelong learning, form relationships with users and act as companions in their daily activities and replacing traditional interfaces such as command line search. Moreover, I expect such agents to be fully integrated with online services, fostering personalisation across platforms.

3 Suggestions for Discussion

Topics for discussion that I think are interesting:

- Deep Neural Networks in place of traditional approaches.
- Long-term interaction with a spoken dialogue system.
- Deep understanding of the user's input.

References

Alexandros Papangelis and Kalliroi Georgila. 2015. Reinforcement learning of multi-issue negotiation dialogue policies. In *SIGDIAL*, Prague, Czech Republic. ACL.

Alexandros Papangelis, Vangelis Karkaletsis, and Fillia Makedon. 2012. Online complex action learning and user state estimation for adaptive dialogue systems. In *ICTAI*.

Alexandros Papangelis, Robert Gatchel, Vangelis Metsis, and Fillia Makedon. 2013. An adaptive dialogue system for assessing post traumatic stress disorder. In *PETRA*, page 49.

Alexandros Papangelis, Robert Gatchel, and Fillia Makedon. 2014a. Assessing and monitoring post-traumatic stress disorder through natural interaction with an adaptive dialogue system. In *JABR*.

Alexandros Papangelis, Ran Zhao, and Justine Cassell. 2014b. Towards a computational architecture of dyadic rapport management for virtual agents. In *Intelligent Virtual Agents*, pages 320–324. Springer.

Konstantinos Tsiakas, Lynette Watts, Cyril Lutterodt, Alexandros Papangelis, Robert Gatchel, Vangelis Karkaletsis, and Fillia Makedon. 2015. A multimodal adaptive dialogue manager for depressive and anxiety disorder screening: A wizard-of-oz experiment.

Zhou Yu, David Gerritsen, Amy Ogan, Alan W Black, and Justine Cassell. 2013. Automatic prediction of friendship via multi-model dyadic features.

Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. 2015. Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness. In *2015 AAAI Spring Symposium Series*.

Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Intelligent Virtual Agents*, pages 514–527. Springer.

Biographical Sketch

Alexandros Papangelis is a Research Engineer at the Speech Technology Group of Toshiba Cambridge Research Lab, working on statistical spoken dialogue. Before that, he was a Post Doctoral Fellow at the ArticuLab at CMU, working in Dialogue Management and Socially-skilled virtual agents. He received his B.Sc. degree from the National and Kapodistrian University of Athens in 2003, in Informatics and Telecommunications, he received his M.Sc. from University College London in 2009, in Machine Learning and his Ph.D. from the University of Texas at Arlington in 2013 and the National Center for Scientific Research "Demokritos", in Adaptive Dialogue Systems for Assistive Living Environments.

Ran Zhao

Language Technologies Institute,
School of Computer Science,
Carnegie Mellon University,
Pittsburgh, PA 15213 USA
rzhao1@cs.cmu.edu
http://www.cs.cmu.edu/~rzhao1/

1 Research Interests

My research interest originated in the study of human-human conversation and interaction. Through leveraging empirical socio-psychological theories and corpuses of human-human communication data, I progressively better understand and discover how humans in dyadic interactions build, maintain and destroy interpersonal relationship.Ultimately, I aim at developing a socially-aware embodied conversational agents that has capability of constructing social bond with human user during the time of assisting him or her to achieve task goals. Specially, my most interested social phenomena in dialogue is rapport. Rapport has been identified as an important function of human interaction.I developed the first computational framework that is able to explain how humans in dyadic interactions build, maintain, and destroy rapport through the use of specific conversational strategies that function to fulfill specific social goals, and that are instantiated in particular verbal and nonverbal behaviors.(Zhao et al., 2014)

1.1 Automatic recognition of conversational strategies

In that work, I focus on automatically recognizing social conversational strategies that in human conversation contribute to building, maintaining or sometimes destroying a budding relationship. These conversational strategies include self-disclosure, reference to shared experience, praise and violation of social norms. By including rich contextual features drawn from verbal, visual and vocal modalities of the speaker and interlocutor in the current and previous turn, I can successfully recognize these dialog phenomena with an accuracy of over 80% and kappa ranging from 60-80%. My findings have been successfully integrated into an end-to-end socially aware dialog system, with implications for virtual agents that can use rapport between user and system to improve task-oriented assistance.(Zhao et al., 2016a)

1.2 Automatically Assessing Dyadic Rapport from Temporal Patterns of Behavior

This work focuses on data-driven discovery of temporally co-occurring and contingent behavioral patterns signaling high and low interpersonal rapport. By utilizing a reciprocal peer tutoring corpus reliably annotated with thin-slice rapport, non-verbals like eye gaze and smiles, conversational strategies like self-disclosure, shared experience, social norm violation, praise and backchannels, I performed fine-grained investigation of how sequences of interlocutor behaviors manifest differently in friends and strangers, uncovering their social intention of facilitating and destroying rapport. I validated the discovered behavioral patterns by predicting rapport against our ground truth via a forecasting model involving two-step fusion of learned temporal associated rules. Our framework performs significantly better than a baseline linear regression method that does not encode temporal information among behavioral features. Implications for the understanding of human behavior and social agent design are discussed.(Zhao et al., 2016b)

1.3 Cognitive-inspired Socially-Aware Hybrid Discourse Planner

The dialogue manager is composed of a task reasoner that focuses on obtaining information to fulfill the user's goals, and a social reasoner that chooses ways of talking that are intended to build rapport in the service of better achieving the user's goals. A task and social history, and a user model, also play a role in dialogue management, but will not be further discussed here. (Matsuyama et al., 2016)

1.3.1 Task Reasoner

The Task Reasoner is predicated on the system maintaining initiative to the extent possible. It is implemented as a finite state machine whose transitions are determined by different kinds of triggering events or conditions such as: user's intents (extracted by the NLU), past and current state of the dialogue (stored by the task history) and other contextual information (e.g., how many sessions the agent has recommended so far). Task Reasoner's output can be either a query to the domain database or a system intent that will serve as input to the Social Reasoner and hence the NLG modules.

1.3.2 Social Reasoner

The Social Reasoner is designed as a network of interacting nodes where decision-making emerges from the dynamics of competence and collaboration relationships among those nodes. That is, it is implemented as a Behavior Network as originally proposed by (Maes,

1989) and extended by (Romero, 2011). Such a network is ideal here as it can efficiently make both short-term decisions (real-time or reactive reasoning) and long-term decisions (deliberative reasoning and planning). The network's structure relies on observations extracted from data-driven models (in this case the collected data referenced above). Each node (behavior) corresponds to a specific conversational strategy (e.g., SD, PR, QE, etc.) and links between nodes denote either inhibitory or excitatory relationships which are labeled as pre-condition and post-condition premises. As pre-conditions, each node defines a set of possible system intents (generated by the Task Reasoner, e.g., "self_introduction", "start_goal_elicitation", etc.), rapport levels (high, medium or low), user conversational strategies (SD, VSN, PR, etc.), visuals (e.g., smile, head nod, eye gaze, etc.), and system's conversational strategy history (e.g., system has performed VSN three times in a row). Post-conditions are the expected user's state (e.g., rapport score increases, user smiles, etc.) after performing the current conversational strategy, and what conversational strategy should be performed next. For instance, when a conversation starts (i.e., during the greeting phase) the most likely sequence of nodes could be: [ASN, SD, PR, SD ... VSN ...] i.e., initially the system establishes a cordial and respectful communication with user (ASN), then it uses SD as an icebreaking strategy, followed by PR to encourage the user to also perform SD. After some interaction, if the rapport level is high, a VSN is performed.

The Social Reasoner is adaptive enough to respond to unexpected user's actions by tailoring a reactive plan that emerges *implicitly* from the forward and backward spreading activation dynamics and as result of tuning the network's parameters which determine reasoner's functionality.

2 Future of Spoken Dialog Research

People pursue multiple conversational goals in dialog (Tracy and Coupland, 1990). Contributions to a conversation can be divided into those that fulfill *propositional* functions, contributing informational content to the dialog; those that fulfill *interactional* functions, managing the conversational interaction; and those that fulfill *interpersonal* functions, managing the relationship between the interlocutors (Cassell and Bickmore, 2003; Fetzer, 2013).There are extensive studies of developing techniques to support human user's propositional goal and interactions goal in dialogue. I think the future of spoken dialogue system should have human-like capabilities that enable to build social bond with the human through generating appropriate behaviors.

3 Suggestions for Discussion

- Evaluation metrics of social dialogue system.
- Cloud-based dialogue system and construct a shared service platform which will enable us work more collaborative.
- Data resource of social dialogue and application of deep learning techniques in dialogue system.

References

Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1-2):89–132.

Anita Fetzer. 2013. 'no thanks': a socio-semiotic approach. *Linguistik Online*, 14(2).

Pattie Maes. 1989. How to do the right thing. *Connection Science*, 1(3):291–323.

Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar J. Romero, Sushma Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *17th Annual SIGdial Meeting on Discourse and Dialogue*.

Oscar J. Romero. 2011. An evolutionary behavioral model for decision making. *Adaptive Behavior*, 19(6):451–475.

Karen Tracy and Nikolas Coupland. 1990. Multiple goals in discourse: An overview of issues. *Journal of Language and Social Psychology*, 9(1-2):1–13.

Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *Intelligent Virtual Agents*, pages 514–527.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016a. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *17th Annual SIGdial Meeting on Discourse and Dialogue*.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016b. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International Conference on Intelligent Virtual Agents*. Springer.

Ran Zhao is a doctoral Student in the Language Technologies Institute at Carnegie Mellon University, where he is supervised by Prof. Justine Cassell. He is currently working on Rapport project, which focuses on developing an Autonomous System for Embodied Conversational Agents(ECAs). His scientific research interests lie in both designing general architecture of multimodal behavior generation and exploring human behavior. Before joining Articulab, he received his B.S. in Computer Science from University of Illinois,Urbana-Champaign working with Prof. Dan Roth in Natural Language Processing and M.S. from Yale University.

Eran Raveh and Iona Gessinger

Saarland University
Computational Linguistics & Phonetics
Saarbrücken, Germany

{raveh|gessinger}@coli.uni-saarland.de

1 Research Interests

Our current research deals with **phonetic convergence** in **human-computer interaction (HCI)**. We are also interested in **intelligent tutoring systems**, especially in the area of **language learning**.

In our current project we are approaching spoken HCI from both the computational and the human perspective, i.e., the user's perspective in the domain of spoken dialogue systems (SDSes).

1.1 Phonetic Convergence

The field of phonetics provides a scientific description for the production of speech. Humans show a high degree of inter-speaker, as well as intra-speaker, variation in their speech productions. In contrast, machines systematically reproduce or reuse speech segments derived from learned patterns using unit selection (Hunt and Black, 1996) or parametric (Zen et al., 2009) speech synthesis, respectively, among others.

In spontaneous conversations, humans tend to phonetically converge to each other, i.e., there is an increase in segmental and suprasegmental similarities between speakers to be observed (Pardo, 2006).

Communication Accommodation Theory (Giles, 1973) postulates that inter-speaker accommodation subserves the function of controlling social distance. Converging to the speech of an interlocutor could thus serve to reduce the social distance.

Apart from the social aspects, phonetic convergence is assumed to support the overall efficiency of spoken interaction by increasing intelligibility and predictability of what is being said.

1.2 Practical Applications

Systems with phonetic awareness and converging capabilities are likely to be more accessible to users than systems that do not accommodate to the speech input. For example, users who are not native speakers of the output language might have a hard time understanding the system's output and would thus benefit from a converging SDS.

Such systems can detect and adapt to variation in the user's speech, potentially making their output more similar to everyday human-human spoken interaction, which may improve the user experience and make the dialogue more fluent and natural.

This kind of adaptation can also be used the other way around, i.e., by *not* converging to the user's speech – perhaps even intentionally diverging from it. This can be effective, for instance, in intelligent tutoring systems for language learning with emphasis on pronunciation, as it is known that pronunciation is usually not emphasized in the language learning process (Grice and Baumann, 2007). Such tutoring systems take advantage of inter-speaker accommodation by giving the learner auditory feedback they can converge to and thus learn to speak in a more native-like manner.

2 Future of Spoken Dialog Research

Multi-system spoken interaction: Nowadays, the most common paradigm for SDSes is a single human interacting with a single machine – e.g., intelligent personal assistance. We believe that one of the next steps is to develop dialogue systems that are also aware of and able to verbally communicate with other machines. This will be handy as soon as speaking machines are more commonly used in everyday interactions and users will be able to dynamically interact with multiple machines to complete a certain task, instead of sequentially completing subtasks with each machine separately. This will also require each system to understand that it was addressed, based on previous utterances and discourse markings.

Tolerance towards speech variation: It is a common phenomenon that users who are not native speakers of American English have more difficulty to be understood by an SDS trained on that particular variety of English. Consequently, those users are forced to adapt their speech to the system. We believe that the system should adapt to the users instead. Therefore, tolerance towards phonetic variation in the speech input needs to be further developed in SDSes. Such tolerance can not only be captured by the learned model, but also by taking language-specific phonetic interferences into account.

Personalized speech output: Personalized speech output already exists in the sense that a device can adapt its underlying model of speech to the user over time. In that way, the speech output of the SDS is refined but does not gain more phonetic flexibility. The above-mentioned tolerance towards speech variation could be translated into linguistic awareness towards speech variation. An SDS that is aware of, and able to produce, phonetic variation could dynamically converge to the speech of different users. This may enhance user experience and make the dialogue more fluent and natural.

3 Suggestions for Discussion

- Objective and subjective evaluation methods for SDSes;
- Development of SDSes for various languages, especially for low-resource languages;
- Improving spoken language understanding (SLU) in SDSes (beyond gap-filling, e.g., using world knowledge or prosodic information).

References

Howard Giles. 1973. Accent mobility: A model and some data. *Anthropological Linguistics* 15(2):87–105.

Martine Grice and Stefan Baumann. 2007. An introduction to intonation – functions and models. In Jürgen Trouvain and Ulrike Gut, editors, *Non-Native Prosody: Phonetic Description and Teaching Practice*, De Gruyter, pages 25–51.

Andrew J. Hunt and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. volume 1, pages 373–376.

Jennifer S. Pardo. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4):2382–2393.

Heiga Zen, Keiichi Tokuda, and Alan W. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51(11):1039–1064.

Biographical Sketches

Eran studied Computational Linguistics at the University of Tübingen, Germany, and at Trinity College Dublin, Ireland. His Bachelor thesis introduced a novel approach for automatically generating customizable recall questions for reading comprehension. He deepened his knowledge of speech processing at a startup company in Dublin, developing a multilingual speech synthesis system, and also while working at the University of Stuttgart.

In real life, he plays the piano and saxophone, and occasionally, basketball.

Eran is currently pursuing a PhD in Dr. Ingmar Steiner's group *Multimodal Speech Processing* at Saarland University.

Iona received a BA degree in Romance Linguistics and Phonetics from the University of Jena, Germany, and an MA degree in Speech Science from the University of Marburg, Germany.

She has worked for the German Language Atlas at the University of Marburg and as a teacher for German as a Foreign Language at Lycée Ronsard, Vendôme, France, and at the Federal University of Rio Grande do Sul, Porto Alegre, Brazil.

Iona is currently pursuing a PhD with Prof. Bernd Möbius at Saarland University.

Acknowledgments

This research is funded by the German Research Council (DFG) in project "Phonetic Convergence in Human-Computer Interaction (CHIC)".

Xiaolong Li

Department of Computer & Information & Science & Engineering,
University of Florida
432 Newell Dr,
Gainesville, FL 32611

xiaolongl@ufl.edu

1 Research Interests

In the dialogue systems research community, there is growing recognition that dialogue systems need to support users in increasingly complex tasks. To move in this direction, dialogue systems must perform natural language understanding within richer and richer contexts, including semantic interpretation of user utterances. My research interests are generally in the area of dialogue understanding, specifically for situated dialogues, which happen in a situated environment. My previous work has focused on reference resolution in situated dialogues, specifically tutorial dialogues for computer science problem solving. I proposed a semantic parser using a conditional random field to parse referring expressions in situated dialogues. I combined semantic parsing results with dialogue history as well as user action history to perform reference resolution in Java programming tutoring sessions. I am currently working on unsupervised approaches for reference resolution in situated dialogues, which solves the problem by employing the temporal relationship between user utterances and user actions in the environment. The goal is to automatically learn the relationship between referring expressions and their referents in the environment.

1.1 Reference Resolution for Situated Dialogues

The content of a situated dialogue is very closely related to the environment in which it happens (Grosz and Sidner, 1986). As dialogue systems move toward assisting users in increasingly complex tasks, these systems must understand users language within the environment of the tasks. To achieve this goal, dialogue systems must perform reference resolution, which involves identifying the referents in the environment that the user refers to (Iida et al., 2010; Liu et al., 2014; Liu and Chai, 2015). Imagine a dialogue system that assists a novice student in solving a programming problem. To understand a question or statement the student poses, such as, Should I use the 2 dimensional array?, the system must link the referring expression the 2 dimensional array to an object in the environment.

Reference resolution in situated dialogue is challenging because of the ambiguity inherent within dialogue utterances and the complexity of the environment. Prior work has leveraged dialogue history and task history information to improve the accuracy of reference resolution (Iida et al., 2010; Iida et al., 2011; Funakoshi et al., 2012). However, these prior approaches have employed relatively simple semantic information from the referring expressions, such as a manually created lexicon, or have operated within an environment with a limited set of pre-defined objects. I have created an approach that addresses those challenges by combining the learned semantic structure of referring expressions with dialogue history into a ranking-based model. I evaluated the new technique on a corpus of human-human tutorial dialogues for computer programming and have shown that it performs significantly better than the previous state of the art (Li & Boyer, In press). An excerpt dialogue is shown in Figure 1.

Tutor:	table = new int[10][5];
Tutor:	that is where they initialize the size of **the 2 dimensional array**
...	...
	[student adds line of code: `arra = new int[s.length()];`
Tutor:	great!
...	...
	[student adds line of code: `new2=Integer.parseInt(parse1);`
Student:	does **my array** look like it is set up correctly now
Tutor:	umm..... in **the for loop**, what should you be storing in **the array**?
Student:	:)

Figure 1: Excerpt of tutorial dialogue illustrating referring expressions refer to referents in the Java code.

1.2 Future Work

For a dialogue system in a situated environment, extra-linguistic information from the environment should also be considered to make proper decisions for the dialogue manager, since the content of the dialogue is closely related to the state of the environment. I plan to further improve the performance of my reference resolution models by incorporating that rich information. Afterward, I plan to investigate the extent to which this higher performing reference resolution module improves outcomes in an end-to-end evaluation of a tutorial dialogue system for introductory computer programming. My study, which will feature my reference resolution model in one version of the system compared to a control condition version of the system, will shed light on the open question of the extent to which student learning and user satisfaction are improved by higher fidelity reference resolution models.

2 Future of Spoken Dialog Research

I believe that there needs to be more research on situated dialogues and dialogue systems that support complex tasks. For several decades, researchers have created effective dialogue systems for a variety of domains. Based on the current state of research work, I think there will be more research on building dialogue systems that can support complex real-world tasks such as problem solving. These kinds of dialogue systems need to employ information from the system's situated environment. There have been several commercial products, such as Amazon Echo, released to work as smart assistants to human users using spoken language. Although these systems are examples of a limited subset of dialogue systems and do not currently extensively employ active awareness of their environments, it is natural for these systems and others (such as assistance robots) to move on to this type of new functionality in the future. To achieve such a goal, dialogue systems must be able to understand user utterances in the situated environment, which requires the systems to leverage multimodal information from the environment.

3 Suggestions for Discussion

- What are the current best practices for building an NLU module for a dialogue system? What are possible ways to improve the NLU module's performance by employing information from the dialogue system's situated environment?

- What are the current best practices to encode simple and complex domain knowledge respectively for a dialogue system? In what ways does the design of the dialogue manager depend upon this encoding?

- What are some of the challenges in making a dialogue system that can learn from the interactions with its users and adapt based on what it has learned?

References

Funakoshi, Kotaro and Nakano, Mikio and Tokunaga, Takenobu and Iida, Ryu. 2012. A Unified Probabilistic Approach to Referring Expressions. *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 237-246.

Grosz, Barbara J and Sidner, Candace L. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 175-204.

Iida, Ryu and Kobayashi, Shumpei and Tokunaga, Takenobu. 2010. Incorporating Extra-linguistic Information into Reference Resolution in Collaborative Task Dialogue. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1259-1267.

Iida, Ryu and Yasuhara, Masaaki and Tokunaga, Takenobu. 2011. Multi-modal Reference Resolution in Situated Dialogue by Integrating Linguistic and Extra-Linguistic Clues. *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, 84-92.

Liu, Changsong and Chai, Joyce Y. 2015. Learning to Mediate Perceptual Differences in Situated Human-Robot Dialogue. *Proceedings of AAAI 2015*, 2288-2294.

Liu, Changsong and She, Lanbo and Fang, Rui and Chai, Joyce Y. 2014. Probabilistic Labeling for Efficient Referential Grounding Based On Collaborative Discourse. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 13-18.

Xiaolong Li, Kristy E. Boyer 2016. Reference Resolution in Situated Dialogue with Learned Semantics. *Proceedings of SIGDial 2016*

Biographical Sketch

Xiaolong Li is a Ph.D. student in the Department of Computer & Information Science & Engineering at the University of Florida. He received his master's degree in Computer Science from Zhejiang University, Hangzhou, China, and his bachelor's degree in Computer Engineering from Northwest Polytechnical University, Xi'an China. His research interests lie in situated dialogue understanding.

David Cohen

Carnegie Mellon University
23, NASA Research Park
Moffett Field, CA 94035

david.cohen@sv.cmu.edu

1 Research Interests

My research interests are in building intelligent computer programs which can relate to people through dialog. I think the most interesting problems related to this goal are in **dialog system architecture**, **evaluation**, and **machine teaching**. I have interest in the application of these programs toward the betterment of humankind, for example through **tutoring systems** and **personal assistants**.

1.1 Dialog System Architecture

1) What type of expertise does it take to build a dialog system? 2) What quality of dialog can you achieve per unit of training data and developer effort? 3) What set of capabilities can you implement per unit of training data and developer effort? These questions are answered if you know the answer to a fourth question: What dialog system architecture are you using?

The answers today are too often: 1) Developers need domain, linguistics, speech recognition, user experience, machine learning, and programming expertise to build a good dialog system. 2) A developer can obtain passable quality for a simple task after acquiring dozens of annotated in-domain dialogs, high task success rate and user satisfaction after collecting big data through a deployment. 3) For every question a system can answer, the developer must hard-code database queries and NLG templates. Clarification and handling of out-of-domain and out-of-capability requests must be done from scratch. Instead of building broad-capability dialog systems, engineers (and marketing teams) end up focusing on getting users to try only what is implemented. The answers should be: 1) Developers require domain expertise. 2) The architecture yields a high-quality dialog system immediately after the domain is encoded. 3) Developers can build a system capable of fluent reasoning and conversation immediately. My previous work on the YODA dialog system architecture [1] aimed to create a dialog system development process closer to this ideal.

YODA comes with generic SLU, DST, DM, and NLG components which support question-answering and command-and-control across many domains. To build a YODA dialog system requires a developer to specify the domain ontology and lexicon, implement non-dialog actions (for command-and-control), and implement the sensor interface. After this is done, they will have a fully-functioning dialog system. Sample dialog systems include an intent-launching personal assistant app reminiscent of Siri, a smart-home dialog system, and a question-answering system about Carnegie Mellon's course schedule. Future work will involve allowing developers to simply build task-oriented dialogs where the system takes the main initiative, and will support automatic language modeling.

[1] http://davidogbodfog.bitbucket.org/yoda/yoda.html

1.2 Evaluation

Most spoken dialog system evaluation has focused on user satisfaction and task completion. The most influential such approach is the PARADISE framework (Walker et al., 1997). PARADISE presumes that maximizing user satisfaction is the purpose of dialog systems, and goes about defining an approach for predicting user satisfaction based on more readily-available performance and cost metrics.

Intelligent computer programs today, including dialog systems, tend to be very narrow in their capabilities and inflexible in their use. Part of the problem is that researchers do not have the tools to measure and report what the capabilities of their systems are, and that breadth and flexibility are not well-defined research goals. My first work to change this was the paper, (Cohen and Lane, 2016). This paper suggests a model and methodology for measuring the breadth and flexibility of a dialog system's capabilities. The approach involves having human evaluators administer a targeted oral exam to a system and provide their subjective views of that system's performance on each test problem. One goal of this work is that authors will augment their reporting with the proposed approach to improve clarity and make more direct progress toward broadly-capable dialog systems.

1.3 Machine Teaching

Machine teaching is defined in (Zhu, 2015) as the inverse problem of machine learning: finding an optimal training set to allow a learning algorithm to efficiently learn a target model. Dialog systems research has made progress toward end-to-end machine learned dialog systems where a data set is collected for the target task, such as in (Serban et al., 2016). If this work is to lead to broadly intelli-

gent dialog systems, researchers will have to extend their focus beyond learning algorithms, and emphasize creating targeted training curricula to allow machine learning algorithms to learn intelligent behavior. An initial focus should be on quantifying the capabilities of machine learned dialog systems as a function of the training data used to create them.

2 Future of Spoken Dialog Research

I believe that dialog system research over the next 5-10 years will require real users and data on a scale not obtainable by academic researchers. Companies like Microsoft, Interactions, Google, and Facebook will show increased willingness to expose their research dialog systems to the public in order to make research progress. Early experiments such as Microsoft's spectacularly failed teen chatbot (Horton, 2016) exposed the shallowness of existing machine teaching and learning approaches, and will be followed up by experiments where researchers think more about the relationship between learning environment and what is learned.

Dialog state tracking will likely be standardized and generalized in the next 5-10 years, allowing task-general dialog state trackers to be trained. Standards will be developed to support more complex dialog states than current systems; for example, instead of assuming dialog state is a single object in a database, dialog state will commonly consist of more complicated data structures.

3 Suggestions for Discussion

- How should a broad-capability, open-domain dialog system be taught?
- How could we go about making tutoring systems with enough content and depth to personalize to and fully occupy a school child?

References

David Cohen and Ian Lane 2016. *An Oral Exam for Measuring a Dialog System's Capabilities.* AAAI Conference on Artificial Intelligence

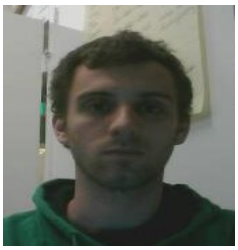
Helena Horton 2016. *Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours.* The Telegraph

Serban, Iulian V., et al. 2016. *Building end-to-end dialogue systems using generative hierarchical neural network models.* AAAI Conference on Artificial Intelligence

Walker, M. A.; Litman, D. J.; Kamm, C. A.; and Abella, A. 1997. *Paradise: A framework for evaluating spoken dialogue agents.* Proc. EACL

Zhu, Jerry 2015. *Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education.* AAAI Conference on Artificial Intelligence

Biographical Sketch



David has received his masters degree from CMU in ECE and his bachelors from UCLA in EE. He was formerly co-founder of Podium, a dialog systems company which specialized in food ordering. He plays basketball and guitar, reads sci-fi and non-fiction, and is an outstanding dancer.

Catharine Oertel

KTH Royal Institute of Technology
Lindstedtsvägen 24
10044 Stockholm
Sweden

catha@kth.se
http://www.speech.kth.se/~cogb

1 Research Interests

My main research interests are in the area of **multimodal multiparty interactions**, and **socially-adaptive dialogue systems** which exploit models derived from such interactions. More specifically, I am interested in measuring the **engagement** of a person in a conversation as well as how this reflect on the conversational dynamics of the whole group of people. I am interested in using **data-driven** approaches which use this knowledge in **human-robot interactions** to increase the collaboration between **children** and improve the perception of attentiveness in **conversational agents**.

1.1 Multiparty Dialogue Systems

In the last years more and more research has focused on the development of multiparty dialogue systems. For such systems it is important to identify whom of several users the dialogue system should address or whether the users are addressing each other and not the dialogue system. It is also important for such a system to know when it should be interrupting a conversation and when it should stay in the background and convey the impression that it is following the conversation, and to what degree etc. These questions are becoming particularly pressing to answer when the purpose of designing the dialogue system is for it to be used in a pedagogical settings or for longterm relationships with its user. In order to be able to answer all these questions, it is necessary to investigate human-human multiparty conversations in terms of their conversational dynamics. In the following paragraphs, I am going to summarize several studies we carried out, and which focused on understanding group dynamics. I will then detail how I am planning to use these findings in a multimodal multiparty dialogue system.

1.2 Conversational Involvement in Multiparty Conversations

For a dialogue system, designed with the purpose of fostering collaboration, it is important to know whether all participants are actually involved in the conversation, and to what degree they are interested in continuing with the conversation. If one of the participants is loosing interest, it can be advantageous for the system to interrupt the current speaker or to change the topic of the conversation. To this end we investigated whether it is possible to both classify the engagement of an individual person as well as the group involvement of the whole group of people (Oertel and Salvi, 2013). In order to make such an estimation as robust as possible for a potential later implementation in an online system, we used "eye–gaze" as a feature. We proposed a number of features (presence, entropy, symmetry and maxgaze) that summarise different aspects of eye-gaze patterns and that allowed us to describe individual as well as group behaviour over time. We used these features to define similarities between the subjects and we compared this information with the engagement rankings, the subjects expressed at the end of each interactions, about themselves and the other participants. We analyzed how these features relate to four classes of group involvement and we build a classifier that is able to distinguish between those classes with 71% of accuracy.

While these results appeared promising, the study did not distinguish between speaker and listener behaviours. In order to fill this gap, we focused on investigating listener categories separately. We distinguished between an "attentive listener", a "side participant" and a "bystander". We then devised a thin-sliced perception test where subjects were asked to assess listener roles and engagement levels in 15-second video-clips taken from a corpus of group interviews. Results showed that humans are usually able to assess silent participant roles. We also found that the frequency of audio backchannel as well as headnods are higher in a "attentive listener" than a "side–participant" and higher in a "side–participant" than a bystander. We also found that mutual–gaze as well as gaze from the speaker are significantly different between the various listener categories (Oertel et al., 2015). In a final study, we then wanted to investigate whether only the frequencies or also the prosodic realizations of the backchannel tokens were different. We therefore used the same corpus to sample backchannels produced under varying conversational dynamics. Amongst other things we wanted to understand i) which prosodic cues are relevant for the perception of varying degrees of attentiveness. We found that duration, intensity and f0 slope are important cues to distinguish between more and less at-

tentive backchannels.

1.3 Socially Aware Dialogue System

Currently I am working on combining the results of both studies by building a "multiparty listener category module" for the dialog system framework IRISTK (Skantze and Al Moubayed, 2012).To also be able to generate the subtleties in audio-backchannels we are currently working on extending the work described in (Oertel et al., 2016), to not only the ranking of different levels of attentiveness in feedback token but also the synthesis. This conversational speech synthesiser will moreover also be able to generate other conversational phenomena such as hesitations, false–starts and self–corrections. We are also working on including convincingly natural synthesis of headnods and smiles.

2 Future of Spoken Dialog Research

I think that in the next 5 to 10 years the field of dialogue research will move even more towards developing multiparty systems as well as systems which can lead a non-task oriented conversation. More and more corpora are becoming available and sensing technology is evolving and becoming more affordable which will foster the possibilities of building such systems. I think that more research will also be devoted to developing dialogue systems designed for longterm interactions. In order to be able to build such systems it will be necessary to investigate how to use information from previous conversation in the current one. For more targeted applications, such as for example pedagogical agents, it will also become more and more important to provide dialogue systems with the capabilities to express empathy and build a rapport.

3 Suggestions for Discussion

- How can we build dialogue systems for longterm interactions ?
- What are the biggest weaknesses for current state-of-the-art dialogue systems?
- How useful can the internet-of-things be for the advancement of dialogue systems ?
- How to make current dialogue systems better suited for interactions with groups of children?

References

Catharine Oertel and Giampiero Salvi. A Gaze-based Method for Relating Group Involvement to Individual Engagement in Multimodal Multiparty Dialogue. 2013. *Proceedings of the 2013 ACM on International Conference on Multimodal Interaction.* 99-106.

Catharine Oertel, Joakim Gustafson and Alan W. Black. Towards Building an Attentive Artificial Listener: On the Perception of Attentiveness in feedback utterances. 2016. *Proceedings of Interspeech 2016.* accepted.

Catharine Oertel, Kenneth A. Funes Mora, Joakim Gustafson and Jean-Marc Odobez. Deciphering the Silent Participant: On the Use of Audio-Visual Cues for the Classification of Listener Categories in Group Discussions. 2015. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction.* 107–114.

Gabriel Skantze and Samer Al Moubayed. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. 2012. *Proceedings of the 2012 ACM on International Conference on Multimodal Interaction.*

Biographical Sketch

Catharine Oertel is a final-year PhD student at the Department of Speech, Music and Hearing at the Royal Institute of Technology in Stockholm, Sweden. She is working on the modeling of multiparty human-human dynamics for building more socially aware dialogue systems within the Horizon 2020 "Baby Robot" Project. She is involved in ongoing research collaboration with both Idiap Research Institute as well as Carnegie Mellon University. She received her M.A. in "Linguistics: Communication, Cognition and Speech Technology"' in 2010, from Bielefeld University, Germany.

José David Lopes

KTH Speech, Music and Hearing
Lindstedsvägen 24
10044 Stockholm
Sweden

jdlopes@kth.se

1 Research Interests

The end goal of my research is to improve communication between humans and between humans and machines. Communication may fail due to a number of reasons: noisy channel, lack of common ground, information about context or even difficulty for one to speak in a different language. While solving all communication problems might be an utopia, my aim is to address smaller problems inherent to one of the reasons above mentioned. For this, I am interested in studying **human-human dialogues** and then apply the outcome of these studies to Spoken Dialogue Systems (SDSs). These studies may show how humans establish common ground and how they handle errors in communication. Therefore they can contribute to the development **SDSs with human-like behavior**. The language barrier is yet another problem that adds to the communication breakdown pile. One of the my long term goals is to create an **SDSs to improve conversation skills** in new language.

1.1 Previous Work

Establishing a common ground in dialogues encompasses several dimensions. During my PhD I have addressed the problem of finding a common ground with respect to lexical items in a dialogue system. The goal was to find a common words that were accepted by both the system and the user to improve the performance of the system. There were two motivations for this. The first one is that there are words that are more easily recognized by a speech recognizer than others. The second one was that there are words that may be preferred by some users and thus the system should try to use them to increase the degree of engagement with the users. To achieve this goal we first studied whether there were words/expressions that users picked up more easily than others (Lopes et al., 2011). In the following step we tried to improve the confidence measure provided by the system to know when a word/expression proposed was not accepted (Lopes et al., 2012). We have built our first rule-based model to automatically choose the ideal word/expression to be used in the system prompt (Lopes et al., 2013). And finally we have also explored a data-driven method to do the same task (Lopes et al., 2015a).

Since I moved to KTH, I started to investigate how to automatically detect communication breakdowns in dialogues with SDSs. Repetitions could be a sign that the communication between the system and the user might have been broken. An automatic method to detect repetitions in interactions with SDSs was developed (Lopes et al., 2015b). This method could be applied to a more general framework to detect miscommunications in dialogues such as the one presented in (Meena et al., 2015). But ultimately, if the goal is to detect the source of the detected miscommunications need to be filtered out by a method such as the one presented in (Georgiladakis et al., 2016). The datasets used in this studies are publicly available (Lopes et al., 2016).

2 Future of Spoken Dialog Research

Where do you think the field of dialogue research will be in 5 to 10 years? The advances in the last few years were rather little, which makes me think that the next coming years will be very exciting. It seems that current solutions for rule-based and data-driven dialogue management have reached their limits. New statistical methods that will be combined with knowledge resources to provide robust and long-term interaction between users and systems.

What do you think this generation of young researchers could accomplish in that time? These generation of researchers should take the opportunity that the field is under the spotlight right now to develop methods that can make spoken dialogue systems something that people use their everyday life.

What kind of questions need to be investigated to get the field to that point? In my opinion, the leading researchers in the field will have to bring new statistical methods into the field that can incorporate information that is today provided either by rule-based models or ad-hoc handcrafted rules.

3 Suggestions for Discussion

- Experiment design and evaluation methods for non-task oriented dialogues and for SDSs for long term interactions.

- Human likeliness in SDSs: Should we aim for systems that behave like humans? Should we use human behavior as a role model that guides the systems we develop? Or should we aim for a different kind of behavior that does not frustrate the user expectations?
- Impact on society: what would be the consequences of the systems will be to the society, what kind of systems should we aim to build, should there be any sort of regulation to limit our research.

References

S. Georgiladakis, G. Athanasopoulou, R. Meena, J. Lopes, A. Chorianopoulou, E. Palogiannidi, E. Iosif, G. Skantze, and A. Potamianos. 2016. Root cause analysis of miscommunication hotspots in spoken dialogue systems. In *Interspeech 2016*, San Francisco, CA, USA, sep.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2011. Towards choosing better primes for spoken dialog systems. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2011, Waikoloa, HI, USA, December 11-15, 2011*, pages 306–311.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2012. Incorporating ASR information in spoken dialog system confidence score. In *Computational Processing of the Portuguese Language - 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings*, pages 403–408.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2013. Automated two-way entrainment to improve spoken dialog system performance. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 8372–8376.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015a. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112.

José Lopes, Giampiero Salvi, Gabriel Skantze, Alberto Abad, Joakim Gustafson, Fernando Baptista, Raveesh Meena, and Isabel Trancoso. 2015b. Detecting repetitions in spoken dialogue systems using phonetic distances. In *Interspeech 2015*.

J. Lopes, A. Chorianopoulou, E. Palogiannidi, H. Moniz, A. Abad, K. Louka, E. Iosif, and A. Potamianos. 2016. The spedial datasets: datasets for spoken dialogue systems analytics. In *10th edition of the Language Resources and Evaluation Conference*, may.

R. Meena, J. Lopes, G. Skantze, and J. Gustafson. 2015. Automatic detection of miscommunication in spoken dialogue systems. In *Proceedings of 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 354–363, Prague, Czech Republic, sep. Association for Computational Linguistics.

Biographical Sketch

José David Lopes is currently a post-doc at the KTH Speech, Music and Hearing department. He got his master from the University of Coimbra in 2008, working on acoustic modelling for noise adverse environments. In the same he joined the Spoken Language System Lab at INESC-ID in Lisbon to start his PhD, which he received form the Technical University of Lisbon in 2013. His thesis was entitled "Lexical entrainment in Spoken Dialogue Systems". After completing his PhD and before starting as a Post-doc at KTH, he worked at the Speech Interactive Research Group at University of the Basque Country, developing a Dialogue System for bus schedule information in the city of Bilbao. After working on the SpeDial european project to develop tools for spoken dialogue system analytics, José's current research is on detecting and developing pauses for Spoken Dialogue Systems.

Carla Gordon

California State University Long Beach
1250 Bellflower Blvd
Long Beach, CA 90840

Carla.Gordon@student.csulb.edu

1 Research Interests

Within the domain of Spoken Dialogue Systems my main research interests lie in **dialogue management**, including developing better solutions for **context dependency**, as well as research related to **Automatic Speech Recognition** systems, and their overall effect on spoken dialogue system functionality. I also have an interest in how dialogue systems can be used to assist in education, specifically language learning.

1.1 Overview of Previous Work

My work in this field has consisted primarily of research related to the development of the dialogue system for the USC Institute for Creative Technology's New Dimensions in Testimony project. This project seeks to preserve the testimony of Holocaust survivors so that future generations may have interactive, real-time conversations with them in a process known as "Time-Offset Interaction." Unlike many dialogue systems which use synthesized voices or text-based responses, this system utilizes actual audio/video recordings of real holocaust survivors.

My initial involvement on this project took place during my undergraduate internship in 2014, during which time it was our goal to discern how many utterances must be recorded in order to support this type of dialogue system. This exact question is the focus of a research paper which I helped author (cf Artstein et al. 2014). The system I helped develop during that time is currently deployed in the Illinois Holocaust Museum and Education Center, as well as the National Holocaust Museum in Washington DC.

This summer (2016) I helped direct a team of interns which conducted data collection for the development of 3 more such systems with 3 new Holocaust survivors. I was also much more involved not only in data collection research, but also in other capacities related to the functionality of the dialogue system itself. I am currently finishing up my internship here at ICT by working with another intern to author a new dialogue policy that would add some context dependency to the system, a feature that is all but absent in the prototype system developed 2 years ago.

2 Future of Spoken Dialog Research

As dialogue systems become more ubiquitous in our daily lives I believe that the field of dialogue research will only continue to proliferate. It is my belief that a multitude of tasks we currently carry out manually will be carried out by interacting with dialogue systems in the future. One such task is word processing. Speech-to-text technology is improving every day, but when you consider all the possible tasks within the realm of word processing it is evident that simple STT translation engines will not be enough for the word processing systems of the future. A fully developed dialogue system would be necessary to accomplish the full battery of tasks associated with word processing such as typographical stylistics, the addition of graphics and charts, and so on.

In the next 5 to 10 years I believe that spoken dialogue systems will come to replace many of the ways in which we interact with technology manually, including word processing, but also in many other situations where we are required to type commands or manually press buttons in order to utilize technology. Just today I watched a presentation by a PhD student here at ICT who helped develop a fledgling dialogue system designed to assist in photo editing. Likewise an intern here this summer developed a dialogue system for a mobile application designed to educate and advise users about the use of sunblock.

It seems, therefore, that one cannot envision the future of technological advancement without the use of dialogue systems to interact with said technology. However, there is clearly a need for more research before we get to this point, specifically I believe in the realm of ASR, because the efficacy of these types of dialogue systems depends so much on properly identifying human utterances.

3 Suggestions for discussion

As it is one of my primary research interests, the topic I would most like to see discussed at the round table is the practical applications of dialogue systems in education and academia. I am greatly interested in how we can design dialogue systems not just to help people accomplish tasks, but also to facilitate education on a wide variety of topics. One of the interns here at ICT this summer was working on just such a dialogue system, designed to help Spanish speakers learn English, and I feel that this is an area of research within the field of dialogue systems that is only just starting to garner attention in the research community.

In addition, and again because it is of primary concern to me, I think a discussion of context dependency within spoken dialogue systems is warranted. I feel that this is a frequently overlooked aspect of dialogue system creation at this point in the evolution of the field, and many of the dialogue systems I've interacted with have a long way to go to be able to properly handle context dependent utterances.

Lastly, I would be very interested in a discussion of how paralinguistic and extralinguistic information could be communicated in spoken dialogue systems, perhaps using virtual humans, but perhaps using a combination of spoken and text based dialogue systems to represent this information. Although I realize this roundtable is specifically about spoken dialogue systems, I also believe that the phenomenon of text-based-dialogue has evolved to such a point as to render it very close to the communicative capacity of spoken dialogue. Therefore, I feel it is not altogether inappropriate to include semiotics as part of a larger discussion about spoken dialogue systems.

References

Ron Artstein, Anton Leusk, Heather Maio, Tomer Mor-Barak, Carla Gordon and David Traum. 2014. *How many utterances are needed to support Time Offset interaction?* Published in the proceedings for the 28th International Florida Artificial Intelligence Research Society Conference.

Biographical Sketch

Carla Gordon is a Linguist by trade (B.A. Linguistics, CSULB 2015), originally interested in phonology and Pronunciation teaching methodology, who has now shifted her research interests to the computational end of the spectrum (M.A. Linguistics with a Special Concentration in Computer Science, CSULB 2018). She has twice interned at the USC Institute for Creative Technologies, (2014, 2016) where she was involved in research pertaining to dialogue system creation the New Dimensions in Testimony project. She will now be joining the ICT staff as a student researcher to continue her contributions to this project..

Outside the realm of dialogue systems, she retains her passion for phonology and pronunciation and has plans to create a new and unique pronunciation teaching application for those who have difficulty distinguishing speech sounds. In her free time (of which there is very little) she enjoys singing and playing guitar, and manages a Youtube channel where she periodically uploads videos of cover songs and the occasional original.

4

5

6

7

8

9

10

11

12

13

14

15

16

Eli Pincus

Institute for Creative Technologies,
University of Southern California
12015 Waterfront Drive, Playa Vista, CA
90094

pincus@ict.usc.edu
http://people.ict.usc.edu/~pincus/

1 Research Interests

Currently, my general research interests lie in methods for **automating** human-human models of communications for use in **embodied spoken dialogue systems**. To this end, most of the research I have conducted involves some or all of the following steps:

1. Analysis of corpora of human-human or human-computer dialogues from a specific domain such as game-play or chit-chat.

2. Implementation of policies (generally in embodied agents) that automate communication patterns found from above analysis.

3. Experimentation to evaluate and iterate on above implementations.

More specifically, I have worked on systems that have been used for studies investigating **user adaptation** , **synthetic voice selection**, **incremental language processing phenomena**, as well as **dialogue selection policies**. The remainder of this section will discuss the main system I have been developing that serves as the test-bed for the main thrust of my graduate research.

1.1 Mr. Clue

"Mr. Clue", is a dialogue agent that can act as clue-giver in a phrase guessing game (Pincus et al., 2014). His design was motivated from analysis of a human-human corpus composed of audio and video recordings of pairs of humans playing a timed word-guessing game (Paetzel et al., 2014; Pincus and Traum, 2014). Evaluations have been conducted with Mr. Clue to test the impact of embodiment, incremental processing, and clue filtering (Pincus and Traum, 2016) on objective game scores and several subjective measures. The system also motivated a novel TTS evaluation (Pincus et al., 2015). So far, results show that many users enjoy playing the game with Mr Clue, but some deficits remain, several of which have motivated current and on-going system improvements. Future work includes implementing an automatic guesser and exploring the effects of adapting the game play style of both agents to certain user attributes.

2 Future of Spoken Dialog Research

In the coming decade I expect dialogue system research to continue to rely increasingly on leveraging technologies from other fields such as computer-vision, electrical engineering, signal processing and others. Some of these technologies will be used to improve the scalability of dialogue systems so that they are capable of acting more natural in real-time conversations. Other technologies will be leveraged to capture more accurate sensor data as well as methods that output interpretations that allow these systems to infer more context of the various situations they find themselves in. Keeping this in mind reminds us that it is imperative that researchers continue to collaborate within and across disciplines in order to move this field forward by achieving the synergies produced from those collaborations.

3 Suggestions for Discussion

Some possible topics for discussion at the round-table:

- Differences and similarities between non-goal oriented vs goal-oriented Dialogue
- User adaptive dialogue systems
- Dialogue systems for extended interaction

References

Maike Paetzel, David Nicolas Racca, and David Devault. 2014. A multimodal corpus of rapid dialogue games. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Eli Pincus and David Traum. 2014. Towards a multimodal taxonomy of dialogue moves for word-guessing games. In *Proc. of the 10th Workshop on Multimodal Corpora (MMC), Reykjavik, Iceland*. Citeseer.

Eli Pincus and David Traum. 2016. Towards automatic identification of effective clues for team word-guessing games. *To appear in Language Resource and Evaluation Conference (LREC)*.

Eli Pincus, David DeVault, and David Traum. 2014. Mr. clue-a virtual agent that can play word-guessing games. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference (AIIDE)*.

Eli Pincus, Kallirroi Georgila, and David Traum. 2015. Which synthetic voice should I choose for an evocative task? In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 105.

Biographical Sketch

Eli Pincus is a PhD student at the University of Southern California advised by Professor David Traum. He is also a graduate research assistant in the Natural Dialogue Group at USC's Institute for Creative Technologies. He has worked as a Lab Associate at Disney Research in the Language Based Character Interaction group and a Research Intern at Nuance Communications in the NLP and AI group. Once upon a time, he was a research assistant in the Spoken Language Processing Group at Columbia University. He holds a master's degree in Applied Mathematics from Columbia.