Proceedings of the 8th Annual

Young Researchers' Roundtable on Spoken Dialogue Systems

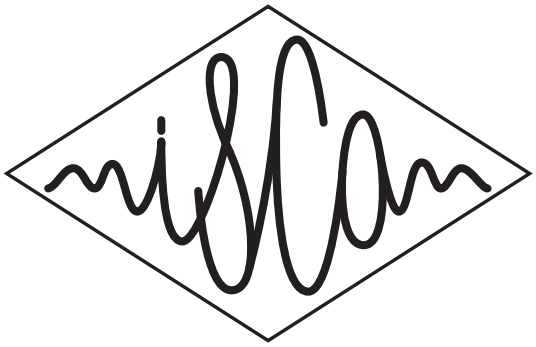
Seoul National University of Seoul, Korea
July 3rd ~ 4th, 2012

http://www.yrrsds.org/

Sponsors

Endorsements

The Association for Computational Linguistics



The International Speech Communication Association

The Special Interest Group on Discourse and Dialogue

Contents

Foreword

Following the success of seven previous workshops in Portland (2011), Tokyo (2010), London (2009), Columbus (2008), Antwerp (2007), Pittsburgh (2006) and Lisbon (2005), we are very happy to welcome you to the Eighth Young Researchers' Roundtable on Spoken Dialogue Systems (YRRSDS 2012) in Seoul, South Korea.

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

This year, we have 25 registered participants, coming from all over the world. The roundtable will also feature 6 guest participants, who will provide us with thought-provoking talks and panel discussions. Alexander Koller (University of Potsdam) and Alex Rudnicky (CMU) are our invited speakers for this year. In addition, Amanda Stent (AT&T) and Jason Williams (Microsoft Research) will tell us more about the exciting work done in their companies. Finally, this year's Industry/Academia panel has an impressive lineup of researchers, including Jason Williams, Antoine Raux (Honda Research Labs), and Luciana Benotti (National University of Cordoba).

This year's roundtable is sponsored by Google, Samsung, Microsoft Research, AT&T and LG, and is also endorsed by ACL, ISCA, and SIGDIAL. We sincerely thank them for their support. We also thank the Seoul National University for providing us with the wonderful venue, and all members of the advisory committee for their insightful comments and suggestions. Thanks also go to our local organizers, Hyuksu Ryu, Kyungduk Kim and WonSeok Choi for their help in deciding the venue, restaurant for the social event, lunch, and so forth. Last but not least, we truly thank this year's participants for their submissions and helpful comments.

We hope you all enjoy this year's roundtable, and have a great time in Seoul!

The YRRSDS organizing committee

Organizing Committee

Timo Baumann
University of Hamburg, Germany

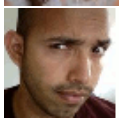
Heather Friedberg
University of Pittsburgh, USA

Jana Götze
KTH Stockholm, Sweden

Srini Janarthanam
HWU Edinburgh, United Kingdom

Pierre Lison
University of Oslo, Norway

Alejandra Lorenzo
Loria, France

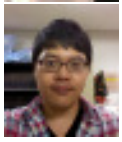


Raveesh Meena
KTH Stockholm, Sweden

Local Organizers

Hyuksu Ryu
Seoul National University, South Korea

Kyungduk Kim
POSTECH, South Korea



WonSeok Choi
University of Sogang, South Korea

Advisory Committee

Seyed Mohammad Ahadi
U Teheran, Iran

Frederic Béchet
U Aix-Marseille, France

Luciana Benotti
U Cordoba, Argentina

Dan Bohus
Microsoft, USA

Rolf Carlson
KTH, Sweden

David DeVault
ICT, USA

Maxine Eskenazi
CMU, USA

Kallirroi Georgila
USC, USA

Peter Heeman
Oregon Health and Science University, USA

Julia Hirschberg
Columbia U, USA

Kristiina Jokinen
U Helsinki, Finland

Tatsuya Kawahara
U Kyoto, Japan

Gary Geunbae Lee
Postech, South Korea

Salam M. Khan
Alabama A&M U, USA

Oliver Lemon
HW Edimburgh, UK

Diane Litman

U Pittsburgh, USA

Wolfgang Minker
U Ulm, Germany

Mikio Nakano
Honda Research, Japan

Rebecca Passonneau
Columbia U, USA

Joelle Pineau
Mc Gill, Canada

Massimo Poesio
U Essex, UK

Antoine Raux
Honda Research, USA

Carolyn P. Rose
CMU, USA

David Schlangen
U Bielefeld, Germany

Stephanie Seneff
MIT, USA

Gabriel Skantze
KTH, Sweden

Amanda Stent
AT&T Labs, USA

David Traum
ICT, USA

Marilyn Walker
UC Santa Cruz, USA

Nigel Ward
U El Paso, USA

Jason Williams
Microsoft Research, USA

Steve Young
Cambridge, UK

Programme

First Day – July 3rd, Tuesday

8:30– 9:00	registration & coffee
9:00– 9:30	welcome and introduction
9:30–11:00	**1st roundtable**
11:00–11:15	discussion of results
11:15–11:30	coffee
11:30–11:45	fanatic poster frenzy
11:45–12:45	poster session, part A
12:45–14:00	lunch (provided)
14:00–15:00	poster session, part B
15:00–15:30	sponsorship talk: **Amanda Stent, AT & T**
15:30–16:30	invited talk: **Alexander Koller**
16:30–16:45	coffee
16:45–18:15	**2nd roundtable**
18:15–18:30	discussion of results
18:30	end of day one
evening	Banquet at the Hoam faculty house

Second Day – July 4th, Wednesday

8:45– 9:00	coffee
9:00–10:30	**3rd roundtable**
10:30–10:45	coffee
11:15–12:15	invited talk: **Alex Rudnicky**
12:15–13:30	lunch
13:30–14:00	sponsorship talk: **Jason Williams, Microsoft Research**
14:00–15:30	special session on tools and toolkits for research in SDS
15:30–16:00	coffee
16:00–17:30	**Industry-Academia panel**
17:30–18:00	conclusion and farewell
18:00	end of day two
evening	self-organized dinner

Invited Guests

Alexander Koller

Invited speaker

Alexander Koller is a professor of theoretical computational linguistics at Potsdam University, Germany. Before joining Potsdam University in 2011 he lead a research group within the Cluster of Excellence in Multimodal Computing and Interaction for three years and where he also received his Ph.D. with a dissertation on *Constraint-based and graph-based resolution of ambiguities in natural language* in 2004 and two M.Sc. in Computer Science and Computational Linguistics. In between, he spent research stays as a post-doc at Columbia University (New York, USA) and at the University of Edinburgh (UK). His research interests encompass areas such as computational semantics, grammar formalisms, and interactive situated NLG, and he is one of the organizers of the GIVE challenge.

Alexander Rudnicky

Invited speaker

Alexander Rudnicky received a B.Sc. in Psychology from McGill University in Montreal in 1975, and a Ph.D. in the same discipline from Carnegie-Mellon University in 1980. After a stay at the University of Toronto, he joined the faculty at Carnegie Mellon, where he is currently a Systems Scientist in the School of Computer Science. In addition to his work in speech recognition, Dr. Rudnicky has studied the influence of language experience on speech perception and visual processes in reading. His current interests include the role of phonology in lexical access, and the design of voice-based interfaces. Dr. Rudnicky is a member of the IEEE.

Amanda Stent

Industry speaker

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

Jason Williams

Industry speaker

As of March 2012, Jason Williams is researcher at Microsoft Research. Before that and since 2006, he was a Principal Member of Technical Staff at AT&T Labs Research. He received a BSE in Electrical Engineering from Princeton University in 1998, and at Cambridge University he received an M Phil in Computer Speech and Language Processing in 1999 and a Ph.D. in Information Engineering in 2006. His main research interests are dialogue management, the design of spoken language systems, and planning under uncertainty. He is currently Editor-in-chief of the IEEE Speech and Language Processing Technical Committee's Newsletter. He is also on the Science Advisory Committee of SIGDIAL, and the board of directors of AVIxD. Prior to entering research, he built commercial spoken dialogue systems for Tellme Networks (now Microsoft), and others.

Luciana Benotti

Panelist

Luciana Benotti is an Associate Professor at the Universidad Nacional de Cordoba (UNC) in Argentina where she is the co-director of the research group Logics, Interaction and Intelligent Systems (LIIS). Luciana finished her Ph.D. in 2010 at INRIA in France, after completing an Erasmus Mundus Masters in Computational Logics. At LIIS, she works on the topics dialogue management, contextual inference and natural language interpretation for dialogue systems. She collaborates with IBM and the Argentinean Research National Agency to develop a framework where end users can develop their own dialogue systems, which is targeted to 3 million high school students. She is a member of the board of the ACL Special Interest group in Semantics of Natural Language.

Antoine Raux

Panelist

Antoine Raux is a scientist at the Honda Research Institute USA, which he joined in Januaray 2009. He obtained a Ph.D. in Language and Information Technologies from Carnegie Mellon University (Pittsburgh, USA) in 2008, a MS in Information Science and Technology from Kyoto University (Japan) in 2002, and an engineering diploma from Ecole Polytechnique (Paris, France) in 1999. During the course of his studies, he performed interships at ATS (Kyoto, Japan), Toshiba R&D (Kawasaki, Japan), and Microsoft Research (Redmond, USA). His research interests lie in all aspects involved in making people interact with machines through speech and other modalities.

Participants

Tadesse Anberbir Awoke
Ajou University
South Korea

Junhwi Choi
Pohang University of Science & Technology
South South Korea

Matthew Henderson
University of Cambridge
United Kingdom

Casey Kennington
University of Bielefeld, Germany

Yonghee Kim
Pohang University of Science & Technology
South Korea

Thanh Cong Le
Dongguk University
South Korea

Injae Lee
Pohang University of Science & Technology
South Korea

Sungjin Lee
Carnegie Mellon University
USA

Changsong Liu
Michigan State University
USA

Teruhisa Misu
NICT
Japan

Timo Baumann
University of Hamburg
Germany

Konstantina Garoufi
University of Potsdam
Germany

Srini Janarthanam
HWU Edinburgh
United Kingdom

Kyungduk Kim
POSTECH
South Korea

Sangjun Koo
Pohang University of Science & Technology
South Korea

Donghyeon Lee
Pohang University of Science & Technology
South Korea

Kyusok Lee
Pohang University of Science & Technology
South Korea

Pierre Lison
University of Oslo
Norway

Alejandra Lorenzo
Loria
France

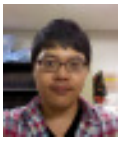
Christopher Mitchell
North Caroline State University
USA

Hyungjong Noh
Pohang University of Science & Technology
South Korea

Aasish Pappu
Carnegie Mellon University
USA

Hyuksu Ryu
Seoul National University
South Korea

WonSeok Choi
University of Sogang
South Korea



Elnaz Nouri
Institute for Creative Technologies
USA

Seonghan Ryu
Pohang University of Science & Technology
South Korea

Ethan Selfridge
Oregon Health and Science University
USA

William Yang Wang
Carnegie Mellon University
USA

Tadesse Anberbir Awoke

Ajou University
Department of Computer Engineering
Ajou University, Suwon, South Korea

email:tadanberbir@gmail.com

1 Research Interests

My research interest is in Natural Language Processing (NLP) and Speech Technology, particularly, I am interested in developing a **Text-to-Speech (TTS)** system for an under-resourced language, Amharic, which is the official language of Ethiopia. My study mainly focuses on **morphological analysis for prosodic assignment** to improve **grapheme-to-phoneme conversion (GTP),** specifically the correct assignment of **geminates** and the insertion of **epenthetic vowels**.

The ultimate goal is to develop a customizable natural sounding free open source TTS system for Amharic language and apply it for different speech enabled applications such as spoken dialogue systems.

1.1 Introduction

Text-to-Speech (TTS) synthesis is a process which artificially produces synthetic speech for various applications. TTS requires the development of efficient Grapheme-to-Phoneme (GTP) converter, which converts a given string of graphemes (letters) into the corresponding string of phonemes (sounds), with proper word boundaries and punctuation marks. However, extracting the correct pronunciation of words and other prosodic features is very challenging.

Amharic is the official language of Ethiopia and belongs to the Semitic language family with the largest number of speakers after Arabic. Amharic uses a unique script, which has originated from ancient language, the Ge'ez alphabet. Amharic writing system is partially phonetic and except redundant sounds, there is more or less a one-to-one correspondence between the sounds and the graphemes. However, the writing system has no way of representing gemination and the correct assignment of geminates and the insertion of epenthetic vowels from grapheme form of a text is challenging.

1.2 Gemination

Gemination is in general a delayed release of a noncontinuant or a prolongation of continuant consonants (Bender and Fulass, 1978). Gemination in Amharic is contrastive and it is one of the most distinctive characteristics of the cadence of the speech, and also carries a very heavy semantic and syntactic functional weight (Bender et al., 1976). Amharic gemination is either lexical or morphological. As a lexical feature it usually cannot be predicted. For instance, ገና may be read as /gəna/, meaning 'still/yet', or /gənna/, meaning 'Christmas'. On the other hand, when gemination is morphological, rather than lexical, it is often possible to predict it from the orthography of the word alone. For example, consider two words derived from the verb root consisting of the consonant sequence sbr 'break', ሰበረው and ይሰበራሉ. The first is unambiguously /sibərəw/ 'break (masc.sing.) it!', the second unambiguously /yissəbbərallu/ 'they are broken'. The fact that the /s/ and /b/ are not geminated in the first word and are both geminated in the second and that the /r/ is geminated in neither word is inferable from the prefix, the suffix, and the pattern of stem vowels. That is, within the verb there is normally some redundancy. Therefore, with knowledge of the lexical and morphological properties of the language, it is possible to predict gemination.

1.3 Epenthesis

Epenthesis is the process of inserting a vowel to break up consonant clusters. Epenthesis, unlike gemination is not contrastive and it is not surprising that it is not indicated in the orthography of Amharic and other languages. But, although it carries no meaning, the Amharic epenthetic vowel /ɨ/ (in Amharic 'ሰርጎ ገባ' (Baye,2008) plays a key role for proper pronunciation of speech and in syllabification.

2 Previous Work

Previously, I worked on the development of Amharic Text-to-Speech system (Anberbir and Takara, 2006) which is a parametric and rule-based system that adopts a cepstral method. The system uses a source filter model for speech production and a Log Magnitude Approximation (LMA) filter as the vocal tract filter. The intelligibility and naturalness of the system was evaluated by word and sentence listening tests and we found promising results.

3 Current Work

Currently, I am working on the NLP part of the AmhTTS system mainly on the Grapheme-to-Phoneme (GTP) conversion using morphological analysis for automatically assigning geminate and epenthetic vowels (Tadesse et al., 2011). In my study, I proposed and integrated a preprocessing morphological analyzer called HornMorpho (Micheal Gasser) into an AmhTTS system. The analyzer takes Amharic text input and outputs Latin transcription marking the location of geminates and epenthetic vowels.

4 Future Plan

Morphological analysis is not sufficient to develop a perfect GTP converter for Amharic TTS and inferring contrastive words such as ገና requires analyzing the context and finding out the parts of speech (POS) of the words. As a future work, I have a plan to develop a complete NLP module and improve the duration modeling using the data obtained from the annotated speech corpus. I have also a plan to collaborate with other researchers and integrate Amharic TTS for Spoken Dialogue systems.

5 Spoken Dialogue Systems (SDS) for Ethiopian Languages

Research on human language technology (HLT) for Ethiopian languages started in the 1990s and there are now a lot of encouraging and valuable works on areas: Machine Translation (MT), Text-to-Speech (TTS), OCR, Morphological Analysis, POS tagging, Stemming, spell checking, Text categorization (Teferra et al.,2011). However, most of these researches are done only for the partial fulfillment of graduation degree and not implemented for real applications.

The lack of language resources, absence of standardization, luck of coordinated research and limited expertise on the HLT areas is also contribute for the poor development of HLT for Ethiopian languages. (Teferra et al.,2011). Till now technologies such as Spoken Dialogue Systems (SDS) are not yet thoroughly explored like other languages and to my knowledge, so far there is no published work on SDS.

So, there is a big demand for researches and application on SDS for Ethiopian languages. Especially with the current advancement of mobile technology (and high number of mobile users) and language processing techniques, spoken e dialog systems will play a great role in everyday life. Therefore, young researchers in the area need to contribute their part by sharing their experience and collaborating with others working on HLT researches.

6 Suggestions for discussion

Possible topics for discussion:

- Using Concept-to-Speech technology to solve the problem of distinguishing geminates and epenthetic vowels.
- Customization: Language independent spoken dialogue systems for under-resourced languages without expensive costs.
- The future of SDS applications for least researched languages in Africa.

References

Baye Yimam. 2008. *የአማርኛ ሰዋሰው (Amharic Grammar).* Addis Ababa.

Michael Gasser, Indiana University: http://www.cs.indiana.edu/~gasser/Research/software.html

M. Lionel Bender, J.D.Bowen, R.L. Cooper and C.A. Ferguson. 1976. *Language in Ethiopia*. London, Oxford University Press.

Solomon Teferra, Binyam Ephrem, Enchalew Yifru, Kassa Tilahun, Lemlem Hagos, Mohammed-hussen and Taye Girma. 2011. *Human Language Technologies for Ethiopian Languages: Challenges and Future Directions* Université Joseph Fourier (UJF) IT PhD Program, Addis Ababa University.

Tadesse Anberbir and Tomio Takara. 2006. *Amharic speech synthesis and its applications to multimedia and telecommunication*. IEICE Technical Report, VOL.105; PAGE.201-206, Japan.

Tadesse Anberbir, Tomio Takara, Michael Gasser, Kim Dong. 2011. M*orphology-Based Grapheme-to-Phoneme Conversion for Amharic Text-to-Speech System*. Proceedings of Conference on Human Language Technology for Development 3-5 May 2011 Bibliotheca Alexandrina Alexandria, Egypt.

Biography

I received my BSc. degree in Biology from Addis Ababa University, Science Faculty in 2000 and M.E. degree in Information Engineering from University of the Ryukyus, Japan, in 2006. Then I joined Addis Ababa University and worked from 2007 to 2008. As of 2008, I am studying my PhD at Department of Computer Engineering, Ajou University, South Korea.

Timo Baumann

University of Hamburg
Department for Informatics
Hamburg, Germany

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

1 Research Interests

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

1.1 Incremental Processing and Evaluation

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

I showed this influence for incremental ASR derived a measure of certainty from the different timing measures and also devised algorithms that improve these incremental properties for iASR using generic filtering mechanisms (Baumann et al., 2009a). Together with my colleagues, we applied the work on evaluation of incremental components to other areas such as semantic interpretation (Heintze et al., 2010), incremental reference resolution (Schlangen et al., 2009), and to n-best processing (Baumann et al., 2009b). As part of our venture into incremental analysis, we built a toolkit to process and visualize incremental data (Malsburg et al., 2009), and the incremental processing toolkit INPROTK (Baumann and Schlangen, 2012b).

1.2 Predictive Processing

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

1.3 Incremental Speech Synthesis

Recently, I have worked on incremental, just-in-time speech synthesis, showing that a system can start speaking with very little utterance-initial processing (Baumann and Schlangen, 2012a) which leads to better system response times and allows for more natural behaviour (Buschmeier et al., 2012). In our approach, synthesis is tightly integrated into the SDS data structures, allowing for seamless, immediate, and on-the-fly adaptation of system utterances.

1.4 Future Work

I plan to further improve the 'conversational' capabilities of speech input, and output for SDS, further working on an integrated architecture for the whole system and including issues such as integrating other modalities and improving prosodic processing for improved naturalness.

2 Future of Spoken Dialogue Research

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

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

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

3 Suggestions for Discussion

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

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

SDSs as tools for language research: In the past, a lot has been learned from dialogue transcripts and later from systematically and on-the-fly altering chat interactions (e. g. with the DiET toolkit). Will it be possible to apply similar alterations to spoken interactions in the near future? If so, what could be learned from manipulated spoken dialogue?

References

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

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

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

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

Timo Baumann, Okko Buß, Michaela Atterer, and David Schlangen. 2009b. Evaluating the Potential Utility of ASR N-Best Lists for Incremental Spoken Dialogue Systems. In *Procs. of Interspeech 2009*, pages 1031–1034, Brighton, UK.

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

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

Silvan Heintze, Timo Baumann, and David Schlangen. 2010. Comparing Local and Sequential Models for Statistical Incremental Natural Language Understanding. In *Procs. of SigDial 2010*, Tokyo, Japan.

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

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental Reference Resolution: The Task, Metrics for Evaluation, and a Bayesian Filtering Model that is Sensitive to Disfluencies. In *Procs. of SigDial 2009*, pages 30–37, London, UK.

Biographical Sketch

Timo Baumann is a researcher and instructor at the University of Hamburg, Germany, and a PhD candidate under the supervision of David Schlangen.

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

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

Junhwi Choi

Pohang University of Science & Technology
Dept. of Computer Science & Engineering
790-784 San 31 Hyoja-dong, Nam-gu, Pohang
Gyungbuk, South Korea

chasunee@postech.ac.kr

1 Research Interests

I have general interests in Natural Language Processing (NLP), especially in error correction for a spoken dialog system. Currently, Spoken Dialog System (SDS) error correction is limited on Spoken Language Understanding (SLU). I focused Automatic Speech Recognition (ASR) error correction for SDS. I already researched error correction for dictation interface. Now, I concerning about how to apply ASR error correction for SDS using my proposed method, seamless error correction. [J.Choi, K.Kim, S.Lee. S.Kim, D.Lee, I.Lee, and G.G.Lee. 2012. ICASSP]

1.1 Voice Only Error Correction for SDS

ASR system is an essential component of SDS. Even when the ASR system has a low error rate, the recognized results frequently include error words. To complete a SDS task perfectly, an error correction process is required. Ignoring ASR errors, re-uttering can be a naïve solution, but unexpected behavior can be occurred. For example, dialog frame slot can be filled by ASR error unexpectedly, so to complete task the dialog frame should be reset. When ASR errors are occurred, if the system knows a current utterance, dialog frame rollback is processed automatically.

The correction process can be performed by selecting an erroneous portion of the text using a keyboard, a mouse, or other devices and speaking replacement text. However, in some usage scenarios, error correction using only voice commands is required. A handicapped person who cannot use either arm may want the error correction to use only voice. In addition, users initially tend to try to correct misrecognized results using their own speech and often remain in the same speech modality even when faced with repeated recognition errors. Therefore, error correction using only voice commands may also be convenient for non-handicapped users.

1.2 Seamless Error Correction

In general, voice-only error correction is a two-step process. In the first step, the users speak a portion of the recognized text to select a target position to correct. Next, the users speak a replacement text. These two steps can perform one correction. However, as McNair and Waibel [1994. ICSLP] suggest, the correction process can instead be performed in a single step. In one-step correction, users speak only their replacement text, and the system automatically recognizes it correctly and finds the error region to replace.

Seamless error correction is processed like one-step error correction, without any explicit command to enter the correction mode. The interface automatically understands the purpose of the utterance whether the intention is to type a new sentence or to correct a misrecognized sentence. Then, the system detects an error region and corrects it. To complement the understanding of user intention, the interface should provide a confirmation process.

The key novel process in the seamless error correction interface is user intention understanding. User intention understanding can be accomplished by the observation of clear speech. User utterances to ASR usually have the characteristics of clear speech, which is a speaking style adopted by a speaker aiming to increase the intelligibility for a listener. To make their speech more intelligible, users will make on-line adjustments; typically, they will speak slowly and loudly, and they will articulate in a more exaggerated manner. Furthermore, the utterances for correction display these characteristics more conspicuously than the utterances for non-correction.

2 Future of Spoken Dialog Research

Focusing on task oriented dialog system, robustness of the system should be emphasized. Not just focusing on the performance of ASR and SLU, error recovery strategy will be important. For ASR error correction, the SDS will automatically recognize occurring of ASR error and find region of error. Then, following dialog would correct it. For SLU error correction, con-

firmation dialog strategy is a solution. The research of naturalness and effectiveness of confirmation dialog strategy will be important.

3 Suggestions for discussion

- When ASR errors are occurred, how to solve dialog frame recovery without ASR error correction?
- The confirmation dialog strategy is suggested for SLU error correction traditionally. Can it be a solution for ASR error correction?

References

Sharon Oviatt and Robert Van Gent. 1996. *Error resolution during multimodal human-computer interaction,* in Proceedings of International Conference on Spoken Language Processing, pp. 204-207.

Christine A. Halverson, Daniel B. Horn, Clare-Marie Karat, and John Karat. 1999. *The beauty of errors: Patterns of error correction in desktop speech systems,* in Proceedings of INTERACT, pp. 133–140.

Arthur E. McNair and Alex Waibel. 1994. *Improving recognizer acceptance through robust, natural speech repair,* in Proceedings of International Conference on Spoken Language Processing.

Seung-Jae Moon and Björn Lindblom. 1994. *Interaction between duration, context, and speaking style in English stressed vowels,* in Journal of the Acoustical Society of America, Volume 96, Issue 1, pp. 40-55.

Rajka Smiljanić and Ann R. Bradlow. 2005. *Production and perception of clear speech in Croatian and English,* in Journal of the Acoustical Society of America, Volume 118, Issue 3, pp. 1677-1688.

J.Choi, K.Kim, S.Lee. S.Kim, D.Lee, I.Lee, and G.G.Lee. 2012. *Seamless error correction interface for voice word processor,* in Proceedings of International Conference on Acoustic, Speech, and Signal Processing.

Biographical Sketch

Junhwi Choi is currently a PhD student at Pohang University of Science and Technology (POSTECH), Korea, supervised by Gary Geunbae Lee. He has general interests in natural language processing. Currently he is focusing on voice only automatic speech recognition error correction.

Konstantina Garoufi

Area of Excellence "Cognitive Sciences"
University of Potsdam
Germany

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

1 Research Interests

How can a computer system generate instructions that will help a human user accomplish their tasks in a large indoor space like an airport or a shopping mall? That is the question my research attempts to provide an answer to. To achieve this, I work on interactive generation of natural-language instructions in situated environments, such as the virtual 3D treasure-hunt game shown in Fig. 1. My interests lie particularly in the intersections of **natural language generation**, **communication in situated environments** and **automated planning**.

1.1 Past Work

The task of generating instructions in 3D environments as in Fig. 1 can involve tackling several different problems, including generation of navigation (e.g. "Go through the doorway in front of you.") or object manipulation (e.g. "Push the left blue button.") instructions, as well as constant execution monitoring and feedback generation (e.g. "No, not that one."). With my colleagues, I have addressed these problems by developing a planning-based approach to language generation (Garoufi and Koller, 2010) that is able to exploit and manipulate the non-linguistic context of communicative scenes besides their linguistic context. By modeling the non-linguistic context in its planning, the system can detect which locations might be convenient for the generation of simple referring expressions that describe to the user the objects they need to identify. This way it can plan ahead and deliberately generate navigation instructions that guide the user to such convenient locations, before finally generating the referring expressions themselves.

Though the above model generates relatively simple and succinct referring expressions, these expressions are not necessarily optimal with respect to effectiveness, i.e., the degree to which they are actually helpful to the user. To address this, we further combined the planning-based approach with a corpus-based measure of effectiveness of referring expressions (Garoufi and Koller, 2011). The system operates by learning a maximum entropy model of referential success from a human instruction-giving corpus we collected (Gargett et al., 2010), and then using the model's weights as costs in a metric planning problem. As a result, it can compute the expressions that

Figure 1: Interactive instruction generation situated in a 3D environment.

are predicted to be the fastest for users to resolve in the given situational contexts. We implemented the system in the framework of the GIVE-2.5 Challenge on Generating Instructions in Virtual Environments[1] (Striegnitz et al., 2011), which is a shared task for the evaluation of natural language generation systems. This evaluation with human users showed that, though not all differences were statistically significant, referring expressions of our system were resolved correctly more often than those of any of the other seven systems participating in the shared task.

1.2 Current and Future Work

The planning-based approach we developed achieves real-time performance in solving non-trivial generation problems. A reason for this is that in this approach we model any perlocutionary effects of communicative actions (i.e., any goals which the actions achieve or are meant to achieve) simply as effects of operators in a planning problem, under the assumption that all these effects will eventually come true as intended (Koller et al., 2010). Clearly, however, communicative actions may fail to have the intended effects, as the user might for instance misunderstand the system's utterances or be uncooperative. It is therefore crucial for an interactive system to constantly observe the user's behavior and estimate to

[1] http://www.give-challenge.org/research

what extent its communicative plans are actually having the intended effects. In ongoing work, we have been experimenting with execution monitoring mechanisms that track the user's eye gaze in order to proactively provide appropriate feedback and enhance the system's communicative success (Staudte et al., 2012; Koller et al., 2012). We anticipate that such eyetracking technology may become mainstream in the not-too-distant future, making it possible for our model to be reimplemented in various different situated communication domains.

Finally, the planning-based approach has the advantage of not being limited to the generation of single noun phrases in isolation, but being capable of also generating entire sentences or discourse. Yet in our work so far we have focused solely on the optimization of referring expressions. As a next step, we aim at the joint optimization of these with other types of utterances such as navigation instructions. A combined measure of effectiveness of referring expressions and navigation instructions could provide the basis for a system that can make interdependent decisions so as to achieve maximization of its overall communicative success.

2 Future of Spoken Dialog Research

I would expect to see everyday-use applications like Apple's Siri getting more widespread in the next years, establishing spoken dialog systems as an inseparable part of our homes, offices and vehicles. Especially mobile devices could become even more powerful, as they can draw from a wealth of data that may not be readily available to other systems (e.g. location, as well as visual and auditory input). Statistical methods and mining from large datasets may continue to play a major role. One of the greatest challenges, however, could be to make such methods aware of the context in which data appears, and model that context in a way that allows system behavior to be tailored to the particular needs that users in specific situations have.

3 Suggestions for Discussion

Topics for discussion I find interesting include:

- **Context awareness.** How can we create systems that are aware of the situational context in which they operate, and are able to interpret non-linguistic signals in parallel with linguistic ones?
- **Optimality.** How can we create systems that learn to make optimal decisions, even when these are not known to system designers? Can such systems generate dialog behavior that outperforms human dialog behavior?
- **Scaling up.** How can we create systems that scale up to large domains?

References

Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Kristina Striegnitz. 2010. The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *Proceedings of the 7th Conference on International Language Resources and Evaluation*, Valletta, Malta.

Konstantina Garoufi and Alexander Koller. 2010. Automated planning for situated natural language generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden.

Konstantina Garoufi and Alexander Koller. 2011. Combining symbolic and corpus-based approaches for the generation of successful referring expressions. In *Proceedings of the 13th European Workshop on Natural Language Generation*, Nancy, France.

Alexander Koller, Andrew Gargett, and Konstantina Garoufi. 2010. A scalable model of planning perlocutionary acts. In *Aspects of Semantics and Pragmatics of Dialogue. SemDial 2010, 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 9–16.

Alexander Koller, Konstantina Garoufi, Maria Staudte, and Matthew Crocker. 2012. Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Seoul, South Korea.

Maria Staudte, Alexander Koller, Konstantina Garoufi, and Matthew Crocker. 2012. Using listener gaze to augment speech generation in a virtual 3D environment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, Sapporo, Japan. To appear.

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

Biographical Sketch

Konstantina Garoufi is currently a doctoral researcher in the Area of Excellence "Cognitive Sciences" of the University of Potsdam. She holds a diploma in Mathematics (University of Athens), a M.Sc. degree in Logic, Algorithms and Computation (same university), and a M.Sc. degree in Language Science and Technology (Saarland University). She has previously worked in the Cluster of Excellence "Multimodal Computing and Interaction" at the Saarland University and in the Ubiquitous Knowledge Processing Lab at the Darmstadt University of Technology. When not navigating through virtual worlds, she enjoys discovering the world of Berlin and its beautiful surroundings.

Matthew Henderson

Department of Engineering,
University of Cambridge,
United Kingdom, CB2 1PZ.

mh521@eng.cam.ac.uk

1 Research Interests

I am interested in **statistical methods** for spoken dialog systems. My research so far has looked at **statistical spoken language understanding**, and is now concerned with applying **structural learning** to the problem of learning Bayesian network user models.

1.1 Robust Discriminative Spoken Language Understanding

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

Mairesse et al. (2009) presented a statistical method for automatically learning a semantic decoder by presenting it as a collection of classification problems, and solving each with an SVM. This was trained and tested on transcribed speech as well as top ASR hypotheses and was found to be competitive with other popular approaches such as hand-crafted grammars. Some work I have done has extended this approach to work more directly on ASR output, in particular confusion networks. It can be shown that by using features extracted from the confusion network it is possible to automatically learn a semantic decoder which is not only much more robust to noise than a handcrafted grammar, but also is more accurate at representing uncertainty in its results (I have submitted a paper on this to SLT).

1.2 Structural Learning of User Models

Key to the POMDP (Partially Observable Markov Decision Process) framework for statistical dialog systems (Thomson and Young, 2010) is the underlying model of the user which comes in the form of a Dynamic Bayesian Network. A partially observed user act and the last machine act serve as inputs to the network, which is updated using Expectation Propagation (Thomson, 2009).

Currently the structure of this network is hand-crafted and highly factored. My current research explores the possibility of automatically learning the structure, which includes dependencies between random variables, the presence of hidden variables, their cardinality and connections etc. A variety of methods have been employed to do structural learning of Bayesian networks (Daly et al., 2011), but this is an extra challenge as the networks are dynamic (they span multiple time slices) and they contain many hidden variables. I am investigating the feasibility of using Expectation Propagation to score candidate model structures.

It is hoped that learning richer networks will allow for modelling users more accurately. For example a simple augmentation to the structure would be a single hidden variable which allows for soft clustering of the user behaviours. Adding more dependencies between sub-goals could model for example the fact that if someone is looking for a restaurant on the riverside, they are more likely to want an expensive place.

1.3 Cambridge Restaurant Information System

One of the systems that we work with in our group is the Cambridge Restaurant Information System, which is a simple dialog system with three slots the user can inform (area, pricerange and food-type) and other slots which can be requested (phone number, signature dish, address and postcode). The current system uses the confusion network decoder described in 1.1, and is available as an online demo at the group website[1].

To demonstrate the system, I have created a multimodal interface to the dialog system which can run online using flash and as an app for both Android and iOS, see Figure 1. This provides multiple views which the user can switch between during a dialog, including a map view and a conversation view. I would be happy to demonstrate this app to anyone interested during the workshop. It is hoped that these applications will facilitate the collection of real data, and allow some research into statistical approaches to multimodal interfaces.

2 Future of Spoken Dialog Research

Spoken dialog systems allow us to interface with computer systems in one of the most natural ways possible, but current slot-filling models for statistical systems can seem inflexible. Slot-filling models are a good first approximation to real dialogs, but in order to allow for more

[1] http://mi.eng.cam.ac.uk/research/dialogue/demo.html

Figure 1: iPad App for Restaurant Information System

natural conversations it might be beneficial to look at alternative models which allow the users to explore the information in new ways. For example in the restaurant domain, such requests as: 'Is it popular with the locals?', 'I want something serving pizza and pasta', 'Is that the cheapest italian there is?' etc. should be possible to answer without having to add a slot in the model for each of them. Achieving this for statistical systems will involve developing new and richer ontologies, semantic decoders and Bayesian user models.

Continued work into using machine learning to learn more of the dialog system statistically will be important over the next decade. In particular, researchers will need to develop methods for automatically learning the ontologies and the structure of user models for larger domains (see 1.2).

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

3 Suggestions for Discussion

- Going beyond slot filling: What sort of models and techniques are necessary for statistical spoken dialog systems to be more flexible in how the information may be accessed by the user?

- Learning larger domains: The challenges of learning both the structure of larger domains, and effective dialog policies in domains of increasing size.

- Dialog systems on your smart phone: Ideas for applications that would sell the idea of dialog systems to the public, and facilitate large scale collection of real data.

References

Rónán Daly, Qiang Shen, and Stuart Aitken. 2011. Learning Bayesian networks: approaches and issues. *The Knowledge Engineering Review*, 26(02):99–157, May.

Nir Friedman. 1997. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 125–133. Morgan Kaufmann.

François Mairesse, Milica Gašić, Filip. Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2009. Spoken language understanding from unaligned data using discriminative classification models.

Blaise Thomson and Steve Young. 2010. Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588, October.

Blaise Thomson. 2009. *Statistical methods for spoken dialogue management*. Phd thesis, University of Cambridge.

Biographical Sketch

After completing his undergraduate degree in Mathematics at Cambridge, Matt Henderson did his master's in Speech & Language Processing at Edinburgh. He is now at the end of the first year of his PhD in Steve Young's group in the department of Engineering, Cambridge. As well as spoken language technology, Matt is interested in language, maths and space.

Srinivasan Janarthanam

Interaction lab
School of Mathematical and Computer Sciences
Heriot Watt University
Edinburgh

sc445@hw.ac.uk
www.macs.hw.ac.uk/~sc445/

1 Research Interests

I am primarily interested **dialogue management**, **user modelling** and **language generation** aspects of spoken dialogue systems.

1.1 Dialogue management

Currently, I am working on a project called SPACEBOOK funded by the European Union. The objective of the project is to build a pedestrian information system with a dialogue interface that helps pedestrian users in navigating and exploring a city. The system will allow pedestrian users can request for navigation instructions to get from A to B, receive information on points of interest (PoI), get amenity information (i.e. cash machines, restaurants), etc.

The pedestrian user will interact with the system using a smartphone app using a headset. The system will present information and receive requests from the user using speech only. This is because, interacting with users visually using the limited mobile screen space and GUI controls may be distract users from performing their primary tasks and also overload them cognitively. Too much speech can also overload users and therefore it is a challenge for the dialogue management module to present information to users effectively.

Complex domains such as these also present a considerable challenge in learning optimal dialogue management policies to manage a variety of tasks (sometimes simultaneously), in optimising language generation choices to present route instructions (using landmarks vs street names, distance vs time, etc) and in optimising dynamic user modelling action choices (adapting to users vs populating the user model). See our demo paper (Janarthanam et al., 2012b) for details. In this project, I am also involved in organising a challenge called the GRUVE Challenge (Generating Routes under Uncertainty in Virtual Environments), in which dialogue/NLG research teams can participate by presenting a dialogue/NLG system that will generate route instructions in a real city-like virtual environments (see (Janarthanam and Lemon, 2011; Janarthanam et al., 2012a) for details).

1.2 Generation of temporal referring expressions

Earlier, I worked in the CLASSiC project (www.classic-project.org), on an appointment scheduling system. In this system, we used reinforcement learning techniques to generate linguistic realisations for appointment slots presented to the users. We evaluated our learned policy both in simulation and with real users and show that this data-driven adaptive policy is a significant improvement over a rule-based adaptive policy, leading to a 24% increase in perceived task completion, while showing a small increase in actual task completion, and a 16% decrease in call duration. This means that dialogues are more efficient and that users are also more confident about the appointment that they have agreed with the system (Janarthanam et al., 2011).

1.3 Dynamic user modelling

For my Ph.D, I worked on a technical support dialogue system that helps users put together a hardware kit for home broadband internet connection. When a dialogue system starts a conversation, it does not always know the user's level of domain knowledge. Users could be experts, intermediates or novices in the domain of conversation. In technical domains, it is essential that the system adapts to the user's domain knowledge levels in order for the conversation to be successful. The objective of my work was to dynamically model the user's domain knowledge (i.e. Broadband Internet and basic computer networking) and adapt the instructions presented to their level of domain knowledge. Using reinforcement learning algorithms, we developed a system that can start a conversation with a user and as the conversation proceeds, it dynamically identifies the user's domain knowledge level and adapt its lexical choice appropriately between jargon and descriptive expressions for the domain objects in the kit. The system learned to optimally switch between information seeking moves (i.e. learn about user's knowledge) and adaptive moves (i.e. use appropriate referring expressions) during the conversations. We showed that adaptive systems built this way produced 99.47% successful task completion and approx. 11% reduction in dialogue duration in comparison to some hand-

coded adaptive systems. See (Janarthanam and Lemon, 2010a; Janarthanam and Lemon, 2010b) for details.

2 Future of Spoken Dialog Research

2.1 Dialogue systems in future

Dialogue systems have started to appear as personal assistants (e.g. Siri and various Android apps such as SpeakToIt). Functionality of such applications, in terms of domain, will increase. Dialogue systems will also emerge as companions to users in order to collect data about their health and well being. They may also support behaviour change by persuading users to for instance, quit smoking, or make more environmentally friendly choices. Future dialogue systems can also be pervasive in the sense that they can be everywhere in several avatars - on your mobile phones, desktops, robot companion, etc.

2.2 Future research in dialogue technologies

Research in dialogue technologies should move towards standardisation in order for rapid development of dialogue interfaces to take place. For example, we need to start using standard dialogue act annotations (see (Bunt et al., 2010)) so that we need not start by defining the dialogue actions for every project. Also, a dialogue systems toolkit built to use the above standard would help young researchers to work on interesting problems in dialogue systems research without worrying about other modules that their work may depend upon.

Another interesting area of work is migration of the agent over several devices (Wallace et al., 2012). A personal assistant agent should be able to migrate from one's mobile phone to a tablet to a desktop. This will let the user to interact with one agent over several devices that he operates. The issues will be as to how to adapt to different devices and their capabilities. Related to this is the issue of forgetting. Long time companions need to be able to forget irrelevant details, compress and archive outdated information, etc much like human memory works in order to be able to serve as natural companions to humans (Lim et al., 2011).

3 Suggestions for Discussion

- Next killer application: Smart mobile assistants, dialogue systems in healthcare and behaviour change domains.
- Standardization of interface definitions between dialogue system modules and rapid development toolkits.
- Evaluation: Global competition, universal metrics for comparing disparate systems.

Biographical Sketch

Srinivasan Janarthanam is currently a research associate at the Interaction Lab, Heriot Watt University at Edinburgh. He received his Ph.D from the University of Edinburgh in 2011. He was a UKIERI (2007-10) scholar funded by the British Council. Previously, he worked as a research associate in Amrita University, India and as an applications developer in iNautix Technologies, India. He has a Masters in Intelligent Systems from the University of Sussex, UK and was a Commonwealth Scholar during this period. He did his undergraduate degree in Computer Science and Engineering from Bharathiyar University, India.

References

H. Bunt, J. Alexandersson, J. Carletta, J. Choe, A. Chengyu, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of LREC 2010, Malta*.

S. Janarthanam and O. Lemon. 2010a. Adaptive Referring Expression Generation in Spoken Dialogue Systems: Evaluation with Real Users. In *Proc. SIGDIAL 2010, Tokyo, Japan*.

S. Janarthanam and O. Lemon. 2010b. Learning to Adapt to Unknown Users: Referring Expression Generation in Spoken Dialogue Systems. In *Proc. ACL 2010, Uppsala, Sweden*.

Srini Janarthanam and Oliver Lemon. 2011. The GRUVE Challenge: Generating Routes under Uncertainty in Virtual Environments. In *Proceedings of ENLG / Generation Challenges*.

S. Janarthanam, H. Hastie, O. Lemon, and X. Liu. 2011. 'The day after the day after tomorrow?' A machine learning approach to adaptive temporal expression generation: training and evaluation with real users. In *Proc. SIGDIAL 2011, Portland, US*.

S. Janarthanam, O. Lemon, and X. Liu. 2012a. A web-based evaluation framework for spatial instruction-giving systems. In *Proc. of ACL 2012, South Korea*.

S. Janarthanam, O. Lemon, X. Liu, P. Bartie, W. Mackaness, T. Dalmas, and J. Goetze. 2012b. Integrating location, visibility, and question-answering in a spoken dialogue system for pedestrian city exploration. In *Proc. of SIGDIAL 2012, South Korea*.

M. Y. Lim, W. Ching, and R. Aylett. 2011. Human-like Memory Retrieval Mechanisms for Social Companions. In *Proc. of AAMAS 2011*.

I. Wallace, M. Kriegel, and R. Aylett. 2012. Migrating Artificial Companions (Demonstration). In *Proc. of AAMAS 2012*.

Casey Kennington

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

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

1 Research Interests

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

1.1 Situational Dialog

In my most recent work (Kennington and Schlangen, 2012), we showed in a small domain that jointly using properties of objects in a shared visual context (color, shape, and spatial relations), the words and linguistic structure of an utterance, as well as knowledge of the previous utterance, were necessary to *understand* what action the user wanted the dialog system to make, what object to take that action on, and what the resulting state of the shared visual world should be. It is clear that words and linguistics, as well as a recognition of objects in the shared space, and previous reference contributed to NLU. This strand of research focuses on **situational dialog**, where the human and the dialog system have visual knowledge about the shared situation. That is, both are part of a scene and can interpret objects in that scene.

1.2 Incrementality

Dialog by definition is **incremental** in that an utterance itself is an incremental unit of a dialog (Schlangen and Skantze, 2009). Further, language unfolds over time (Frazier, 1987) in incremental units which are on a finer-grained scale than sentences, or even words. Humans can perceive and understand an utterance on-line as it is being spoken. This is already a motivation for building dialog systems that can process input incrementally, but it is also a matter of practicality; a dialog system that continually processes new input will provide a more natural user experience. My work until this point has emphasized *incremental* natural language understanding. I currently use, and will continue to use, the Inpro Toolkit (Baumann and Schlangen, 2012), which is an implementation of an incremental dialog framework.

1.3 Future Work

Obtaining more information from the context such as eye gaze, gestural information (following (Bergmann et al., 2011)), and information from the speech signal, are next on my research agenda, whether or not these add to the incremental understanding of the dialog, and how much. This, of course, requires more **interaction**, which is a real testbed for dialog systems. My previous work has focused on interpretation only, but adding interaction and more information sources will force more use of natural language generation, which is a longer-term future research goal.

2 Future of Spoken Dialog Research

- Dialog systems that can interact more naturally (not relying on fragmented turn-taking), thus causing less frustration to the human user. This requires better overall understanding from all aspects of dialog. The less frustrated human users are with a dialog-system, the more likely they are to use dialog systems, which means more usable data for research in all areas of dialog. Some of our research focuses on batch processing, not so much direct interaction,

which is important, but there needs to be a balance. As we improve our dialog systems, we always need to take time and see how things fit into a real interaction scenario.

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

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

3 Suggestions for Discussion

- *What is interaction?* A dialog system can appear to *act* human-like but if understanding doesn't take place, is interaction really taking place? (Edlund et al., 2008) gives a nice overview of methods of which one could infer that "humanness" in and of itself is useful and worth pursuing in dialog research. I agree, but what role does understanding take?

- *What makes up* NLU*?* What parts of dialog need to be jointly predicted? Which parts can be modularized? For me, when it comes to NLU, it is important to know what is fundamental and what is an appendage. I believe that it begins, of course, with the utterance, the words and some linguistic understanding, as well as what objects are referred to in the real world. The manner of speech or other cues, though very important, are appendages to that.

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

- *SDS Evaluation:* In many areas of computational linguistics, there are automatic ways of evaluating the performance of a system. This is useful in many areas, but in discourse and dialog I must object to an automated evaluation (for now). During my masters work I spent a lot of time in machine translation. One reason I left that field is because the standard evaluation method has caused the entire field to try to improve MT on the basis of how well it performs against a metric that is known to have weaknesses. In contrast, papers which involve dialog almost always have to come up with a unique way of evaluating their module or system which, in my opinion, is how it should be. It will allow evaluation to evolve over the years, forcing us to take a critical look at how our systems are evaluated, and the field will eventually come to an accepted, de facto way of proper evaluation.

References

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

Kirsten Bergmann, Hannes Rieser, and Stefan Kopp. 2011. Regulating Dialogue with Gestures — Towards an Empirically Grounded Simulation with Conversational Agents. In *Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 88–97. Association for Computational Linguistics.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8–9):630–645.

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

Casey Kennington and David Schlangen. 2012. Markov Logic Networks for Situated Incremental Natural Language Understanding. In *Proceedings of SIGdial 2012*, Seoul, Korea. Association for Computational Linguistics.

David Schlangen and Gabriel Skantze. 2009. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, number April, pages 710–718, Athens, Greece. Association for Computational Linguistics.

Biographical Sketch

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

Yonghee Kim

Pohang University of Science and Technology
San 31, Hyoja-Dong,
Pohang, 790-784, Korea

ttti07@postech.ac.kr

1 Research Interests

My research interest is concentrated on the **user adaptive chatting system** which makes the conversation more attractive and leads much closer machine-human interaction. There's various ways to achieve this, and among them, my first approach is about **user modeling** for **long-term memory** support. As **emotion recognition and expression** are essentials for establishing rapport, they are also not-to-be-missed subjects. Since voice is more informative than plain text, I hope there are plenty of opportunities to improve technology.

1.1 Past and current work

As the first step of my research, I have co-worked for developing and improving chatting system based on lexico-syntactic patterns and named entity information. The system basically tries to find an example sentence in DB which matches with user input, then selects the paired system utterance as a system output. Using lexico-syntactic patterns, the system can improve the coverage of pattern matching dramatically. Also it strengthens the robustness of the system against erroneous speech recognition result. The chatting system with erroneous ASR result is my main portion of the overall work.

1.2 Future work

As I mentioned in previous part, I will add user model to the existing system. The system would be able to recognize and store some information related to the user. After that, the system remarks it in right time or applies it to other tasks. In the previous researches, these kinds of systems not only memorize the information, but also expect user's demand and suggest what the user would like to do. However, most of them are designed to work within specific domain like web searching. Extracting user-related information and modeling it might be arising issues in chatting system.

Another research topic is sentiment analysis and emotion expression. Usually chatting language is consists of very short expressions with specific tone and speed depending on the speaker's purpose or mood. The word itself gives very few information. Without context, we cannot understand the proper meaning in many cases. Multimodal features are needed in addition to plain text input. Although the state-of-the-art already achieves a lot of thing, there still remain many points to be explored.

All these works are necessary for improving user experience of human-computer interaction. The system would seem alive rather than a machine and people can accept it as a reliable talker.

2 Future of Spoken Dialog Research

Existing systems have focused on the surficial features of conversation and imitate it. They works very well in well-defined domain. Future systems will extend to deeper layer of conversation. User models and ontologies reflecting some parts of human thinking process will get importance in this research area. They will be able to accomplish more complex jobs and effectively reduce the human efforts.

More efficient method of building and maintaining existing system will be developed continuously. Technologies like unsupervised learning reduces the cost of system build, and it will be helpful to process huge amount of (corpus) data.

Sentiment analysis and emotion expression are challenging problem to deal with. However, even with very simple emotional expression (like agree or disagree expression), the conversation gets much more natural. Basic emotional system will be adopted soon, in the industry.

3 Suggestions for discussion

- Data mining from user utterances: Effective detection and extraction of user related information (named entities).
- User model for chatting system: How to handle and apply the user specific data.
- Evaluation: Global competition, universal metrics for comparing disparate chatting systems. The metrics should reflect user satisfaction, propriety of the response, etc.
- Sentiment detection via prosody analysis: The speaker's intention is represented by the tone, speed, rhythmic pattern and even more various features. We can distinguish the valid features from noncritical ones.
- Emotion expression: The system can control their utterance to convey its intention effectively.

References

Takahiro Matsumoto, Satoru Satake, Takayuki Kanda, Michita Imai and Norihiro Hagita. 2012. Do you remember that shop?: computational model of spatial memory for shopping companion robots, HRI '12 Proceedings.

Seong-Yong Koo Kiru Park, Hyun Kim, and Dong-Soo Kwon. 2011. A dual-layer user model based cognitive system for user-adaptive service robots. 2011 IEEE International Conference.

Adams, P.H. and Martell, C.H. 2008. Topic Detection and Extraction in Chat, 2008 IEEE International Conference.

Panagiotis Giannoulis and Gerasimos Potamianos 2008. A Hierarchical Approach with Feature Selection for Emotion Recognition from Speech, 2012 IREC Conference.

Biographical Sketch

Yonghee Kim is currently a Ph.D student in the CSE Dept. of POSTECH, Korea, under the advisor of Gary Geunbae Lee. He is doing research on chatting system, especially user modeling for long term memory support.

He was born in Seoul, Korea, and his bachelor's degree is also earned from POSTECH. He loves classical music, and enjoys writing his daily thought or ideas in his place.

Sangjun Koo

Intelligent Software Laboratory
Dept. of Computer Science & Engineering
Pohang University of Science & Technology
Pohang. South Korea

giantpanda@postech.ac.kr

1 Research Interests

The main stream of my research topics focuses on dealing with **multimodal model in computer-assisted language learning (CALL)**. Recently, the need of learning system for foreign language (especially for English) has emerged greatly. Although there exist a number of assistant system in the fields, most of them only operate on verbal and written input from user. Considering the problem, I have been working on developing integrated CALL system with multimodal interface including haptic device and gesture recognizer. What I expect from the methodology is users' immersion over the system, which would consequently leads to better educational performance.

With construcing the integrated CALL system, **Referring expression(RE) resolution and generation in multimodal system** has become the major research issue. Since the system would have to deal with scenarios using REs for educational purpose, it would be necessary to devise and apply effective RE processing alogithm. I have been searching for the probablistic model which could deliver proper mechanism in processing REs.

1.1 Integrated system for CALL

The main motivation of the system is to enhance POMY, the virual enviornment with dialog management system. Main purpose of the system was to provide immerse environment to Korean students who want to boost their English skill (Noh et al. 2011). Several tasks are given to learners such as path finding, market in the system. Learners should make conversation with NPC(non-playable character) in order to acquire necessary information to achieve given tasks. This conversation consists of spoken utterances which are made by learner and system.

Since the system could communicate with learners and provide them new type of English education framework successfully, the possibility to enhance immersity with various scenarios based on multimodal medium. In other words, system can communicate with learner not only by means of spoken words but also by means of vibrating haptic devices or gesture recognizers. One can assume the task of buying commodities, for instances, where learner has to buy specified items. The system then gives recommendation by emphasizing certain objects on screen and learner may select appropriate items with hand gestures. If a learner make mistake in selecting correct object or make grammar errors in dialog, system warns learner by making a vibration.

Each task scenarios can be considered as single module in overall CALL system structure. In order to expect much immerse responses of learners, the structured take form of game. The game simulates daily life of student and learners have to achieve given tasks in order to get high score. Ultimate goal of project is to provide advanced environment for English learners and to get experimental intution in boosting dialog system.

1.2 Referring expression resolution and generation

Referring expression(RE) is the expression which *distinguishes specific referred objects* from other objects. What makes RE interesting is that the distinct properties of referred objects are not given intrinsically but are given through speakers/writers' intention. Although Dale pointed that the mechansim of generation and understanding of RE follows Gricean Maxims (Dale and Reiter 1995), it is still a challenge to generate proper model for RE.

There exist previous works over processing REs. Kelleher et al. suggest incremental algorithm to resolve and generate REs (Kelleher and Kruijff 2006). The algorithm generates and resolves REs by extracting distinct features of each objects iteratively. Funakoshi et al. (Funakoshi et al. 2012) suggest the method which establish bayesian network for REs. They suppose that each observed words in single RE are decided by concepts denoted by the word, referents of the RE and presupposed referring domain and consider them as random variables. REs are able to be resolved by calculating emergence probabilities of each referent in given bayesian model structured with presented variables.

The aim of my recent research is to apply several algorithm for REs to integrated system which I have mentioned in previous subsection since processing REs can help to generate certain scenarios using gestural deixis and other referring devices. Furthermore, it would help to figure out how referring expressions can help language learning in given system. In other words, it is expected

to find more applicable RE model through the process which result in performance improvement in dialog system.

2 Future of Spoken Dialog Research

The probabilistic dialog management technique has been developed quite further enough to be used on real world circumstances. Nowadays, there exist a number of spoken dialog systems which are used to assist both general and specific tasks.

However, it seems that the essence of probabilistic model tends to rule out potential advantage of using rule-based NLP model which is constructed with syntactic and semantic rules in language. It is implied that humans are likely to have innate ability to handle concepts which can be presented as certain rules. It would be probably different from how probabilistic model handles them.

So, I believe that researchers would probably be able to establish revised model which would give better performance over resolving the meaning of given expressions in near future.

To achieve the goal, cooperation work among computational linguistics and cognitive science would be highly recommended. This effort would help dialog systems to boost their performance which eventually leads to emergence of applications in real world in much general forms.

3 Suggestions for Discussion

- It is necessary to design adequate architecture in order to build successful multimodal spoken dialog system for language learning. Which components would be essential to build multimodal system and which scenarios would be possibly adopted in given system? In addition, how can we evaluate the performance of certain type of dialog system?

- Are there any ideas of general method to describe semantic and pragmatic properties over certain dialog? Although the question is considered as ideal case, further discussion would help to have more intuition over the problem.

- Parallelism have become global trend in Computer Science field. The parallel approaches have potential to improve computational performance which would result in delivering better solution or reducing computation time. I am sure that there would be an opportunity to discuss the topic of applying parallel methodology on spoken dialog system. Any creative idea of parallel application would be appreciated.

References

R. Dale and E. Reiter. Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions. *Cognitive Science*, 19(2):233–263, 1995.

K. Funakoshi, M. Nakano, T. Tokunaga, and R. Iida. A unified probabilistic approach to referring expressions. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 237–246, Seoul, South Korea, July 2012. Association for Computational Linguistics.

J. D. Kelleher and G.-J. M. Kruijff. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 1041–1048, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

H. Noh, K. Lee, S. Lee, and G. G. Lee. Pomy: a conversational virtual environment for language learning in postech. In *Proceedings of the SIGDIAL 2011 Conference*, SIGDIAL '11, pages 344–346, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-10-7.

Biographical Sketch

Sangjun Koo is currently a MS student at the Intelligent software laboratory in Pohang University of Science and Technology. He received his B.S. in Computer Science and Engineering with Combined Minor of Language Technology from Seoul National University. Currently he is working on BRL(Basic Research Lab) integration project.

His interests include reading linguistic books, walking around the campus and thinking and complaining about new semantic and pragmatic model over and over.

Thanh Cong Le

Dongguk University
Faculty of Computer Science
Gyeongju Campus
Korea

thanhlct@gmail.com

1 Research Interests

In general, I am interested in **Artificial Intelligence (AI)**, **Natural Language Processing (NLP)**, **Machine Learning**, **Statistical**, **Probability**, and especially in **Dialogue Systems**. I do not focus on full **Spoken Dialogue System (SDS)**, but only on the **Speech Understanding** and **Dialogue Management** modules, two important subparts of a SDS. These modules are traditionally implemented with **Hidden Vector State Models (HVS)** and **Partially Observable Markov Decision Processes (POMDP)**, in combination with learning methods such as **Supervised Learning**, **Unsupervised Learning** and **Reinforcement Learning**. Concretely, I am studying state of the art methods and trying to implement them. Moreover, I would like to propose improvements and new methods.

1.1 Previous and Current Work

Some years ago, I focused on different technologies and AI in general. I worked on Chess game programming with **Alpha-beta pruning** algorithm, and on **Indoor Positioning** with methods like **K-NN**, **Kalman Filter** and **Particle Filter**. Moreover, I worked in Interactive Video Player, IPTV and Video on Demand (VOD).

From last year, I started to concentrate in NLP in general, but my final goal is to implement really smart SDSs. I already implemented many approaches to solve different basic NLP problems, such as **Language Models**, **Parsing** using different methods, **Part of Speech Tagging (POS Tagging)**, **Information Retrieval**, **Question and Answering** and so on. Moreover, I implemented many methods of **Supervised/Unsupervised Learning** such as **Regression**, **Neural Network**, **Support Vector Machine (SVM)**, **K-Means** and in particular, I did a small work in **Recommender System** for Videos. But always keeping the focus on my passion, Dialogue Systems, by reading textbooks, papers, designs etc.

My current work consists in exploring SDSs. In particular, I plan to implement two subparts of SDS, namely Semantic Processing and Dialogue Management. Then, I would like to build a full SDS on basic level.

1.2 Future Work

As mentioned above, my final goal is to build a really smart Dialogue System. This is also desired by many AI researchers, who try to build a human-like intelligent system. In my case, I concentrate on the area of Dialogue Systems.

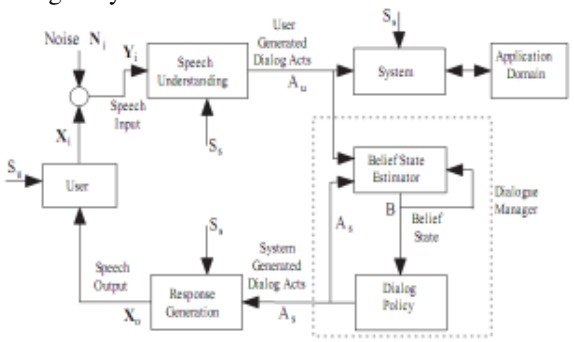


Figure 1. Statistical Model of Spoken Dialogue (Young, 2002)

As proposed in (Young, 2002) a **Statistical Model of a SDS** is shown in Figure 1. [1]The system operates cyclically. It begins with a default, system-initiated, dialogue act A_s which converted to an acoustic signal X_o inviting the user to speak. Based on the current user state, S_u, the user generates a signal X_i which is corrupted by noise before being input to a speech understanding component as the acoustic Y_i. The noisy speech signal Y_i is decoded to give a set of dialogue acts A_u. These dialogue acts are interpreted causing the system state S_s to be updated. The system's next action depends on the belief state B. This belief state encodes the information in the user's state S_u and the system state S_s needed to take the appropriate next action. There will usually be uncertainty in this estimation, and hence in the general case, B will not be a single discrete variable, rather it will be a probability distribution over the combined event spaces of S_s and S_u. Based on the system's new updated belief state B, a new system dialogue act A_s is generated and the cycle repeats.

[1] The description taken from (Young, 2002)

We can see that the two most important parts in such a SDS are the Speech Understanding and Dialogue Management modules which purposes are, respectively, understand what user said, and know what the system should answer to user. More precisely, they are the Semantic Processing, Dialogue Act Detection and Dialogue Management modules, where the methods applied to implement them are HVS, **Tree-Augmented Naïve Bayes Networks** and POMDP.

My future work focus on building SDSs able to perform automaton retrieval and learning, and to refine semantic processing and dialogue management. In other words, that is the **Dialogue Systems as Babies**.

It is well known that parents teach their babies by repeating utterances and actions, and increasing the complexity of the teaching over time. Babies learn from the images, sounds and further information they receive based on senses as vision, hearing, touch etc. This learning is continuous, and happens even while they sleep. Even if they learn some concept which is not correct, they can still correct it given they are taught again.

Finally, we have huge and diverse information over Internet which a computer could explore. The problem is how to create a system able to automatically use this information to refine its own knowledge, helped by users who could correct the mistakes done by the system. In the SDS I am currently working on, I try to do this by investigating the Speech Understanding and Dialogue Management modules.

2 Future of SDS Research

Many researchers would like to build Dialogue Systems that behave as a human being with burning passion. So, we will have Dialogues Systems like that in future. But not in the near time.

Every day we get new achievements, take us closer to the goal. Especially helpful are the young researchers, pushed by their passion, their bold ideas, their thirsts, and their wishes to improve the achieved results over time. However, young researchers usually take a long time to relearn old achievements, since they are not able to get guide or help from experienced researchers.

Spoken Dialog System, in future, will focus on **Statistical Model**, and Machine Learning methods, especially in **Reinforcement Learning**. Moreover, the systems need to have **Data Mining** ability, refine its own knowledge from huge information resources over Internet, and be able to apply natural knowledge with more powerful ways for **Meaning Representation, Semantic Representation,** and **Context-Aware**.

3 Suggestions for discussion

- Survey of SDSs: These methods, models and designs applied in SDSs. Comparing each of them and giving concrete examples for helping young researchers.
- Achieved Systems related to Dialogue Systems: Given clear achievements. Young researchers are able to keep on improving them. Such as Apple Siri, CLASSiC, MIT's TINA, WASTON, Wolfram Alpha and so on.
- Hidden Vector State models, Partially Observable Markov Decision Processes: These best methods for implementing SDS.
- Machine Learning: Learning is the key for really smart Dialogue Systems.

References

David Griol, Lluis F. Hurtado, Encarana Segarra, Emilio Sanchis. 2008. A Statistical Approach to Spoken Dialog System Design and Evaluation. Speech Communiation.

Jason D. Williams, Steve Young. 2006. Partially Observable Markov Decision Processes for Spoken Dialog System. Computer Speech and Language.

Oliver Lemon, Oliver Pietquin. 2007. Machine Learning for Spoken Dialogue Systems. Interspeech.

Oliver Lemon. 2008. Adaptive Natural Language Generation in Dialogue using Reinforcement Learning. SEMdial.

Steve Young. 2002. The Statistical Approach to the Design of Spoken Dialogue Systems. Cambridge, England.

Yulan He, Steve Young. 2004. Semantic Processing Using the Hidden Vector State Model. Computer Speech and Language.

Biographical Sketch

Thanh Cong Le is a current Master student at Faculty of Computer Science, Dongguk University, Korea. He got B.S degree at Faculty of Information Technology, Hungyen University of Technology and Education, Vietnam, 2010. He is interested in Dialogue System, Machine Learning, and Data Mining. He is currently studying in his ideal about real intelligent Spoken Dialogue System, named is "**Dialogue Systems as Babies**".

Donghyeon Lee

Pohang University of Science & Technology
Dept. of Computer Science & Engineering
790-784 San 31 Hyoja-dong Nam-gu Pohang
Gyungbuk South Korea

semko@postech.ac.kr

1 Research Interests

A dialog corpus is an essential resource for developing data-driven spoken dialog system (SDS). The dialog corpus should be labeled with semantic tags to train models for SDS. However, preparing an annotated dialog corpus requires laborious and time consuming tasks. Developing and maintaining SDS from the annotated corpus is a tedious tasks because SDS has many components such as automatic speech recognition (ASR), spoken language understanding (SLU), and dialog management (DM). How do we reduce tedious human efforts for these tasks?

I have focused on reducing the laborious work for SDS. There are two main aspects of my research work. One is the **automatic annotation** of the raw dialog corpus by using **unsupervised approach**. The other aspect is the **implementing toolkits** to support the development and management of data-driven SDS.

1.1 Unsupervised Approach to Semantic Annotation

The semantic annotation process consists of two steps; designing and labeling step. In designing step, a linguistic expert defines types of labels and generates a guideline for annotators. In labeling step, the annotators label semantic tag referring to the guideline. It is difficult to determine types of labels in the guideline. Moreover, labeling is labor intensive and time-consuming.

Although active and semi-supervised approaches can reduce human labor of traditional supervised approaches, it still relies on annotated corpora with a human intervention and cannot reduce time for the designing process. In contrast, unsupervised clustering approach does not require annotated corpora. The limitation of the classical unsupervised clustering method such as K-means algorithm is that the number of cluster is fixed a priori. It is particularly difficult to find the adequate number of clusters in a dialog system setting. It is not always possible to know the number of clusters in advance. A human analysis is required to determine the number of clusters. For a fully unsupervised approach this number should be automatically set.

To address this problem, we used hierarchical Dirichlet process Hidden Markov Model (HDP-HMM) which is a Bayesian non-parametric approach that infers the effective number of clusters. Our model basically combines the HDP-HMM with a content model for a dialog corpus. We also include the two following Bayesian extensions to improve the model: an entity model and a background model.

We can measure the clustering performance by using the manually annotated sets as the target clustering. This strategy is not the best way to evaluate the clustering results in two reasons. First, we cannot guarantee that the human annotation is the best answer. Determining the types and labeling are ambiguous tasks for humans. Second, the ultimate goal is to use the automatically labeled dialog corpus for a dialog system. Thus, it is necessary to evaluate performance using the dialog system for better clustering evaluation. In our experiments, we used example-based DM method (Lee et al., 2009; Kim et al. 2010), which is one of the data-driven dialog modeling techniques.

We conducted the evaluation of dialog system to evaluate the effectiveness of our model in dialog system development. In our experiment, we showed that our unsupervised model achieves a competitive result in comparison to a system using manually annotated corpus (Lee et al., 2012).

1.2 Dialog System Development Toolkit

Developing data-driven SDS involves two main problems: first, the integration of the various components and second, the corpus preparation. To provide efficient and a convenient development environment, a workbench tool for data-driven SDS called DialogStudio has been developed, which satisfies the following criteria.

- Well-designed workflow
- Simple and easy corpus annotation
- Language synchronization
- Domain and methodology neutral workbench tool
- Easy training and testing environment

There are five steps in DialogStudio workbench tasks: design, annotation, language synchronization, training and running. The workflow considers semantic, dialog, and knowledge tasks. For synchronizing each component, we proposed the concepts of language synchronization.

The usability of DialogStudio was validated by developing dialog systems in three different domains with two different dialog management methods (EBDM and POMDP). The results of evaluation showed that using DialogStudio is effective in developing and maintaining data-driven SDS (Jeong et al., 2008; Jeong et al. 2011).

2 Future of Spoken Dialog Research

Recently, SDSs have been widely applied to user interfaces in many devices such as television, mobile phone and tablets. Thus, SDSs can be used by many people for general domain instead of specific domain. In such environment, I believe that an adaptability of SDS would be an essential property.

By growing network capabilities, many service providing systems are implemented under client-server architecture. In this architecture, very large log data can be collected in servers. The log data are useful resource to improve models for SDS. In addition, we can apply external resources generated from on-line conversation and social networking. In the future, it becomes more important for SDS development and management to minimize a human intervention by learning from the big data for the adaptability.

We tried to apply the concept of the daydreaming to SDS. Daydreaming is performed by a self-evolutionary process, which involves analyzing the log data, trying to extract the patterns, and updating the models to learn from the log data (Lee et al., 2010).

3 Suggestions for discussion

- Adaptation: how to analyze the log data and learn from the log data?
- User modeling: how to model user's behaviors and apply them for SDS?
- Evaluation methods: Real user evaluation is costly. What are the best methods for simulated user evaluation?

References

Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, Gary Geunbae Lee. 2009. *Example-based dialog modeling for practical multi-domain dialog system*. Speech Communications, 51(5): 466-484.

Donghyeon Lee, Kyungduk Kim, Cheongjae Lee, Junhwi Choi, Gary Geunbae Lee. 2010. *D3 Toolkit: A Development Toolkit for Daydreaming Spoken Dialog System*. Proceedings of the 2nd International Workshop on Spoken Dialog Systems Techonology (IWSDS 2010).

Donghyeon Lee, Minwoo Jeong, Kyungduk Kim, Gary Geunbae Lee. 2012. *Unsupervised modeling of user actions in dialog corpus*. Proceedings of the IEEE international conference on acoustics, speech, and signal processing (ICASSP 2012).

Kyungduk Kim, Cheongjae Lee, Donghyun Lee, Junhwi Choi, Sangkeun Jung and Gary Geunbae Lee. 2010. *Modeling Confirmations for Example-based Dialog Management*. Proceedings of the 2010 IEEE Workshop on Spoken Language Technology (SLT 2010).

Sangkeun Jung, Cheongjae Lee, Seokhwan Kim, Gary Geunbae Lee. 2008. *DialogStudio: A workbench for data-driven spoken dialog system development and management*. Speech Communications, 50 (8-9):683-697.

Sangkeun Jung, Cheongjae Lee, Kyungduk Kim, Donghyeon Lee, Gary Geunbae Lee. 2011. *Hybrid User Intention Modeling to Diversify Dialog Simulations*. Computer Speech and Language, 25(2): 307-326.

Biographical Sketch

Donghyeon Lee received the B.S. degree in Computer Science Engineering from Sungkyunkwan University in 2006. He is currently a Ph. D student of Pohang University of Science and Technology. His research interests include unsupervised SLU and dialog system development toolkit. In his free time, Donghyeon likes watching sports.

Injae Lee

Pohang University of Science & Technology
Dept. of Computer Science & Engineering
790-784 San 31 Hyoja-dong Nam-gu Pohang
Gyungbuk, South Korea

lij1984@postech.ac.kr

1 Research Interests

My research interests include the **domain selection** in multi-domain dialog systems. More specifically, I am interested in understanding the current user intention and its dialog history. I also am interested in detecting out-of-domain utterances in multi-domain dialog systems.

1.1 Current work

The goal of our research is to select the domain of the domain expert which generates the system utterance in response to the user utterance in a multi-domain dialog system. The user intention should be considered carefully to achieve the goal. The user intention is considered by analyzing both the current user utterance and its dialog history.

In our approach, we listed and ordered the whole domains according to linguistic, semantic, and keyword features extracted from the user utterance. To reflect the semantic feature, both the generic spoken language understanding (SLU) result which is domain independent information and the domain specific SLU result are considered. We used the features to calculate the domain suitability in each domain and then each domain is listed by descending order for the domain selection efficiency (Cheongjae Lee et al., 2009).

We, then, applied 'in-domain verifier' and decided the final domain (Mikio Nakano et al., 2009). We used the existence of the contents from the database as the essential feature of the in-domain verifier. The existence of the contents indicates that the user utterance was analyzed and processed in the acceptable domain.

To retrieve contents from database considering the dialog history, we accumulated named entities (NEs) in each domain. Accumulated NE slots are used to retrieve contents from the database even when we did not extract from the current user utterance. In case of no content result from the database search for the accumulated slots, we only used the current slots and retrieved the contents again.

1.2 Future work

The dialog history should be considered when there is no content database result in all domains. We should verify the reason for no contents retrieval whether the domain is selected by mistake or there is no content in the correct domain. Identifying the correct NE will help to choose the correct domain without neither contents database nor the dictionary in each domain since collecting all contents information is not realistic.

2 Future of Spoken Dialog Research

Many researchers focus on improving dialog system performance and generating the appropriate system action given the user input or reducing human efforts. I think dialog systems will have been developed and the systems will become convenient, easy and natural enough for people to use. Lots of software using dialog interfaces are developed and supplied to the practical field.

I think the next focus of the spoken dialog system will be emotion extraction and emotion expression. The dialog system will be able to be used for personal chatting, medical diagnosis, and counseling. The knowledge of psychology or sociology will be applied to the system, then, systems will become widely used by people.

3 Suggestions for discussion

Possible topics for discussion:

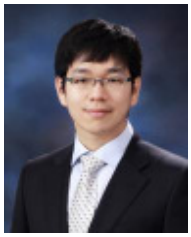
- The efficient and effective way of real world knowledge.
- How to attract people to use spoken dialog interface in many applications.
- What do you expect to the spoken dialog system?

References

Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. 2009. Example-based dialog modeling for practical multi-domain dialog system.

Mikio Nakano, Shun Sato, Kazunori Komatani, Kyoko Matsuyama, Kotaro Funakoshi, and Hiroshi G. Okuno. 2011. A Two-Stage Domain Selection Framework for Extensible Multi-Domain Spoken Dialogue Systems.

Biographical Sketch



Injae Lee is currently a M.S. student in Computer Science Engineering in Pohang University of Science and Technology since 2010. He received the B.S. degree in Computer Science Engineering from Hanyang University. His research interests include spoken dialog system, multi-domain dialog system, and domain spotting. His hobbies include reading, and taking photos.

Kyusong Lee

Pohang University of Science & Technology
Dept. of Computer Science & Engineering
790-784 San 31 Hyoja-dong Nam-gu Pohang
Gyungbuk South Korea

kyusonglee@postech.ac.kr
home.postech.ac.kr/~kyusonglee

1 Research Interests

My current research interest is computer-assisted language learning (CALL) in a virtual game-like environment. One-on-one tutoring is a much more effective teaching strategy than group and is conceivably the best way to improve speaking ability in conversational English. However, due to the high cost and limited number of native English speaking teachers, among other factors, many learners of English as a Foreign Language (EFL) usually have limited opportunities to speak English in a natural language learning setting. Therefore, in Korea despite spending enormous time and energy in learning English, many students still have difficulties with communicating in English. For these reasons, there is considerable interest in research regarding English education in Korea. Robot learning and E-learning have been garnered interest in future second language learning education (Lee, et al 2010). Our research group is investigating in more interesting, motivating, economical and efficient ways to learn English. Game-based learning is currently being evaluated as an educational method because of its benefits of high user motivation, attention, and interest. Moreover, the concept of social community is gaining popularity so that many people are spending time interacting in a virtual environment with their own characters

1.1 Language Learning Game

We are developing Dialog-based language learning game. It is an educational game designed for language learners to convey interactive conversations with in-game characters in interactive immersive environments, such as post office, library, shops, on the street, etc. Students have meaningful interactions with non-player characters (NPCs) to complete tasks in each game mission. For the domains that students were exposed to, we selected such domains as path-finding, market, post office, library, and movie theater to ensure having students practice conversations in everyday life setting.

1.2 Related Work

Several systems have been developed for the purpose of language teaching and learning in interactive environments. The Tactical Language and Culture Training System (TLCTS) is one of the most successful systems. It targets members of the U.S. military who need to acquire basic communicative skills in Arabic and knowledge of cultural differences in the given zone of operations like Iraq (Johnson et al., 2007). The Spoken Electronic Language Learning (SPELL) (Morton & Jack, 2005) provides opportunities for learning languages in functional situations such as going to a restaurant and expressing likes and dislikes. Its key element has been developed to recognize grammatical errors, especially those made by non-native speakers. Recast feedback is provided if the learner's response is semantically correct but includes grammatical errors. This system combines semantic interpretation and error checking in the speech recognition process. Thus, it uses special speech recognition to identify and respond to both correct and erroneous speech. DEAL, a Spoken Dialog System developed at KTH (Kungliga Tekniska högskolan, Royal Institute of Technology) focuses more on creating entertaining gameplay (Brusket et al., 2007). DEAL which uses the trade domain, specifically a flea market situation, provides hints about things the user might try to say if he or she is having difficulties remembering names of items, or if the conversation has stalled for other reasons.

1.3 Conversation Assistance (Hint Generation)

Students often do not know the proper responses to advance in game encounters. But when provided with answers directly, student feel bored when playing the game. Therefore, instead of revealing the answer, we give the hints in order for students to speak properly on their own. The Ranking-based DM makes it possible to generate the hints based on the most probable user utterances (Noh, et al 2011). N-best results of user utterances are used for Hint generation. To increase the variety in student interactions with the system, we define the several different ways to show the hints. 1) Comprehension question - A list of potential answers

with distractors 2) Grammar question - the correct sentence with embedded grammatical errors to be corrected.

1.4 Field Study

we developed a Spoken Dialog-Based Language Learning Game (DB-LLG), called Postech Immersive English Study (Pomy). To investigate the educational effects of our approaches using the educational game, Pomy, our research group performed a field study at a Korean elementary school. A course was designed in which students had meaningful interactions with NPCs in an immersive virtual environment to verify the cognitive effects of our approaches. The result showed significant difference in the listening, vocabulary, and speaking skills. The results showed that our CALL approaches can be an enjoyable and fruitful activity for students (Lee et al., 2011).

2 Future of Spoken Dialog Research

Nowadays, many commercial mobile applications using spoken dialog systems are launched such as Apple Siri, Samsung S-Voice, etc. In 5 to 10 years, many others applications will be released such as game, health care, and educational system. Therefore, people will expect more human-like dialog systems. It would be more natural if each NPC in game has different personality when interacting with players. If the performance of dialog system is getting stable, dialog systems embodied with personality and emotions would be attractive to customers and researchers in the near future.

3 Suggestions for discussion

- Evaluation in Educational SDS
- Interacting with virtual and robotic agents.
- SDS for Educational applications and game.

References

Brusk, J., Wik, P. and Hjalmarsson, A. (2007) DEAL: A Serious Game for CALL Practicing Conversational Skills in the Trade Domain. In: The Proceedings of SlaTE-Workshop on Speech and Language Technology in Education. Pennsylvania, USA.

Johnson, W.L. Wang, N. & Wu, S. (2007) Experience with serious games for learning foreign languages and cultures. Proceedings of SimTecT, Australia, 2007.

Lee, K., Kweon, S., Lee, S., Noh, H., Lee, J., Lee J., Kim, H., Lee, G. (2011). Effects of language learning game on Korean elementary school students. Proceedings of the SLaTE 2011.

Lee, S., Noh, H., Lee, J., Lee, K., Lee, G. G., Sagong, S., & Kim, M. (2011) "On the Effectiveness of Robot-Assisted Language Learning", ReCALL Journal, 23(1), 25-58.

Morton, H. & Jack, M. A. (2005) "Scenario-based spoken interaction with virtual agents", Computer Assisted Language Learning, 18, 171-191.

Noh, H., Lee, S., Kim, K., Lee, K. & Lee, G.G. (2011) "Ranking Dialog Acts using Discourse Coherence Indicator for Language Tutoring Dialog Systems", Proceedings of the 3rd International Workshop on Spoken Dialogue Systems Technology (IWSDS), Granada.

Biographical Sketch

Kyusong Lee received the B.S. degree in Computer Science Engineering from Soongsil University in 2010. He is currently a Ph.D. student in computer Science Engineering from Pohang University of Science and Technology. His research interests include grammar question generation, grammatical error detection, intelligent computer-assisted language learning, spoken dialog system, machine learning. In his free time, Kyusong likes travelling, cycling, and running.

Sungjin Lee

Language Technologies Institute,
Carnegie Mellon University,
Pittsburgh, Pennsylvania, USA

sungjin.lee@cs.cmu.edu

1 Research Interests

My research focuses on improving statistical dialog modeling, especially via unsupervised machine learning techniques to minimize human intervention. I am also interested in its use in attractive applications such as foreign language education.

1.1 Past work

Although computer-based English learning is in the center of interest, this method usually fails to provide the opportunity for free conversation and stays at the level of simple repetition of the given text. These teaching-learning methods cannot provide persistent motivation for learners to reach the high proficiency levels in foreign languages. Considering the shortcomings for the current teaching-learning methodology, I have been investigating English learning systems using spoken dialog technology in immersion context based on the assumptions of second language acquisition theory and practice.

For the past several years, I have been participating in developing robots and virtual agents as educational assistants. The robots, called Mero and Engkey, were designed with expressive faces, and have typical face recognition and speech functions allowing learners to have a more realistic and active context (Lee et al., 2011a). The virtual environment, called Pomy, presents a virtual reality immersion, where learners experience the visual, aural and tactual senses to help them develop into independent learners and increase their memory and concentration abilities to a greatest extent (Noh et al., 2011). These systems can perceive the utterances of learners, especially Korean learners of English. Korean learners' production of the sound is different from those of native speakers, resulting in numerous pronunciation errors. Therefore, our research group has collected a Korean-English corpus to train acoustic models. In addition, since language learners commit numerous grammatical errors, we should consider this to understand their utterances. Thus, we statistically infer the actual learners' intention by taking not only the utterance itself but also the dialog context into consideration, as human tutors do (Lee et al., 2010).

While free conversation is invaluable to the acquisition process, it is not sufficient for learners to fully develop their L2 proficiency. Corrective feedback to learners' grammatical errors is necessary for improving accuracy in their interlanguage. For this purpose, I investigated grammatical error simulation and detection methods which play key roles in helping learners to use more appropriate words and expressions during a conversation. When a learner produced ungrammatical utterances, our system provides both implicit and explicit negative and positive feedback in a form of elicitation or recast, which was manifested as effective ways in the second language acquisition processes. To provide corrective feedback on grammatical errors, we use a method which consists of two sub-models: the grammaticality-checking model and the error-type classification model (Lee et al., 2011b). Firstly, we automatically generate grammatical errors that learners usually commit (Lee et al., 2011c), and construct error patterns based on the articulated errors. Then the grammaticality-checking model classifies the recognized user speech based on the similarity between the error patterns and the recognition result using confidence scores. After that, the error-type classification model chooses the error type based on the most similar error pattern and the error frequency extracted from a learner corpus.

1.2 Current and future work

My current research interests lie in improving statistical dialog modeling, especially via unsupervised machine learning techniques.

The motivation for this comes from the fact that many systems are presently moving from being simple lab simulations to actual deployed systems with real users. These systems furnish a constant flow of new data that needs to be processed in some way. Our goal is to minimize human intervention in processing this data. Previously, data had to be hand-annotated, a slow and costly process. Recently crowdsourcing has made annotation faster and less expensive, but all of the data still has to be processed and time must be spent in cre-

ating the annotation interface and tasks, and in quality control. In an effort to minimize the level of human intervention, I have developed a fully unsupervised approach to user simulation in order to automatically furnish updates and assessments of a deployed spoken dialog system (Lee and Eskenazi, 2012a). Also, I have proposed the use of unsupervised approaches to improve components of partition-based belief tracking systems. The proposed method adopts a dynamic Bayesian network to learn the user action model directly from a machine-transcribed dialog corpus. It also addresses confidence score calibration to improve the observation model in an unsupervised manner using dialog-level grounding information (Lee and Eskenazi, 2012b).

My future research topics include real-time incremental dialog strategy learning and adaptation. This kind of research is expected to greatly improve many limitations caused by the use of user simulators for dialog strategy learning. Also, it can open up the door to self-evolving systems.

2 Future of Spoken Dialog Research

Most spoken dialog systems have been designed towards information seeking, sometimes degenerated to voice search tasks. I believe that, with the advances in speech and language technology, spoken dialog systems would appear in many areas of our daily life, such as foreign language tutoring, voice-controlled digital devices, and conversational robots for elderly people. To facilitate the development of a wide range of applications, the most important prerequisite is to decouple the domain independent communicative behavior from the task-level properties so that the large dimension of state space can be greatly reduced and the communicative model may be used across all different real domains. Further, it becomes more important for conversational agents to recognize and track users' emotional state. In pursuing this goal, we need to move focus to designing a holistic model considering cognitive and affective aspects together rather than just optimizing task-level efficiency.

3 Suggestions for discussion

- Approaches to user modeling to provide responses tailored to user profiles.
- Approaches to on-line dialog system learning
- Incorporating affective computing into applications of spoken dialog system.

References

Hyungjong Noh, Kyusong Lee, Sungjin Lee, Gary Geunbae Lee. 2011. *POMY: a conversational virtual environment for language learning in POSTECH,* 12th SIGDIAL, Portland, USA.

Sungjin Lee, Hyungjong Noh, Jonghoon Lee, Kyusong Lee, Gary Geunbae Lee, Seongdae Sagong, Moonsang Kim. 2011a. *On the Effectiveness of Robot-Assisted Language Learning,* ReCALL Journal, Vol.23(1).

Sungjin Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee, 2011b. *Grammatical Error Detection for Corrective Feedback Provision in Oral Conversations,* 25th AAAI, San Francisco, USA.

Sungjin Lee, Jonghoon Lee, Hyungjong Noh, Kyusong Lee, Gary Geunbae Lee. 2011c. *Grammatical Error Simulation for Computer-Assisted Language Learning,* Knowledge-Based Systems.

Sungjin Lee, Cheongjae Lee, Jonghoon Lee, Hyungjong Noh, Gary Geunbae Lee. 2010. *Intention-based Corrective Feedback Generation using Context-aware Model,* CSEDU, Valencia, Spain.

Sungjin Lee and Maxine Eskenazi. 2012a. *An Unsupervised Approach to User Simulation: toward Self-Improving Dialog Systems,* 13th SIGDIAL, Seoul, South Korea.

Sungjin Lee and Maxine Eskenazi. 2012b. *Exploiting Machine-Transcribed Dialog Corpus to Improve Multiple Dialog States Tracking Methods,* 13th SIGDIAL, Seoul, South Korea.

Biographical Sketch

Sungjin Lee received the B.S. and Ph.D degree in Computer Science Engineering from Pohang University of Science and Technology in 2006 and 2012. He is currently a postdoctoral researcher of Carnegie Mellon university. His research interests include statistical dialog modeling, machine learning, intelligent computer-assisted language learning, and grammatical error simulation and detection. In his free time, Sungjin likes travelling, reading books, watching movies and spending weekends with his family.

Pierre Lison

Language Technology Group
Department of Informatics
University of Oslo, Norway

plison@ifi.uio.no
http://folk.uio.no/plison

1 Research Interests

My general research interests revolve around **dialogue management**, and more specifically adaptive dialogue management for rich, open-ended domains. As part of my PhD, I am developing an hybrid approach to dialogue management which seeks to combine the benefits of both symbolic and statistical approaches in a single, unified framework. The key idea is to devise compact probabilistic models which take advantage of the internal structure of dialogue, and are therefore easier to learn and generalise to unseen data. I hope to be able to demonstrate that such framework is able to model interaction phenomena which cannot currently be well captured by classical approaches to dialogue management. I also want to show how dialogue management can be applied beyond the traditional slot-filling applications which have received most attention in the field so far.

1.1 Probabilistic rules

I am currently working on a new approach to the design of probabilistic models of dialogue, based on the concept of *probabilistic rules*. Probabilistic rules are essentially structured mappings between input and output state variables, intuitively described in terms of *if... then...else* procedures. The rules provide a compact, uniform encoding for the various models used in a dialogue architecture (for spoken dialogue understanding, management and generation). In essence, they provide the system designers with a powerful specification language to encode their prior knowledge about the problem structure, and therefore yield probabilistic models which are both easier to learn and generalise better than classical, unstructured models. I've given a short description of this framework in my most recent paper (Lison, 2012), which I'm going to present at SIGDIAL this year.

In practice, these rules are used as *templates* for the generation of a Bayesian Network which is then used to perform probabilistic inference, based on standard algorithms. I'm demonstrating in the above-mentioned paper how to perform Bayesian parameter estimation on the parameter of these rules, and applied this technique to the problem of policy learning on a limited data set collected via Wizard-of-Oz experiments.

This framework is currently being implemented in a Java-based open-source platform called **openDial**, which aims to be a generic architecture for the development of spoken dialogue systems. I'm currently developing this platform and testing it for various scenarios related to human-robot interaction, using the Nao robot.

1.2 Plans for Future Work

I'm currently working on extending my approach in several directions:

- The first line of work is to extend the parameter estimation to Bayesian model-based reinforcement learning. The parameter estimation currently operates in a supervised learning mode, which requires expert data. Alternatively, one could estimate the model parameters in a fully online fashion, without any supervisory input, by incorporating model uncertainty into the inference and continuously adapting the parameter distribution from real or simulated interaction experience (Ross et al., 2011). Such learning procedure could be applied both to the estimation of models related to dialogue understanding and interpretation, as well as action selection models. For the latter, we would have to combine Bayesian learning with online planning techniques (Ross et al., 2008).

- Another research direction relates to the extension of the belief update algorithms towards incrementality (Schlangen et al., 2010). We believe that the framework presented in this paper is particularly well suited to perform incremental processing, since the chain of related hypotheses is explicitly captured in the conditional dependencies of the Bayesian Network. A probability change in one initial hypothesis (e.g. the user utterance) will therefore be directly reflected in all hypotheses depending on it (e.g. the corresponding user intention). Extending the belief update algorithm to run incrementally while remaining tractable is however a non-trivial task.

- Finally, the framework which I am currently developing is not confined to dialogue policy but can be used to structure any probabilistic model, from dialogue understanding and interpretation to dialogue management and to natural language generation. It

is therefore possible to use probabilistic rules as a unifying framework for all dialogue models defined in a given architecture, and exploit it to perform *joint optimisations* of dialogue understanding, action selection and generation.

2 Suggestions for Discussion

Possible topics for discussion:

- *Dialogue systems for long-term interaction*: most currently deployed dialogue systems are tailored for short-term interactions, typically lasting from a few seconds to a few minutes at most. How can we design dialogue systems able to handle longer types of interactions? Such ability might be crucial for the development of personal assistants or social companions interacting with one or several users on a regular basis over several weeks or months.

- *Prior knowledge in adaptive dialogue systems*: most theoretical frameworks for dialogue management either assume policies which are either fully designed (i.e. handcrafted by the system designer) or fully learned (using reinforcement learning). But there has not been much work so far on policies combining both learned and designed aspects, apart from a few exceptions such as (Williams, 2008). How do we reconcile learning & adaptation with the exploitation of prior knowledge?

- *Joint optimisations for dialogue processing*: Most spoken dialogue systems are based on a pipeline architecture where each module is designed and optimised in isolation from each other. Pipeline architectures offer some advantages in terms of software integration, but also significant drawbacks when it comes to evaluating the overall quality of the interaction. An alternative approach, outlined in recent work such as (Lemon, 2011) is to *jointly* optimise some of the system components. Can we go even further and come up with optimisation techniques able to optimise parameters for an end-to-end spoken dialogue system, starting from speech recognition all the way to speech synthesis?

References

O. Lemon. 2011. Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation. *Computer Speech & Language*, 25:210–221.

Pierre Lison, Carsten Ehrler, and Geert-Jan M. Kruijff. 2010. Belief modelling for situation awareness in human-robot interaction. In *Proceedings of the 19th International Symposium on Robot and Human Interactive Communication (RO-MAN 2010)*, Viareggio, Italy.

Pierre Lison. 2009. Robust processing of situated spoken dialogue. In Christian Chiarcos, Richard Eckart de Castilho, and Manfred Stede, editors, *Von der Form zur Bedeutung: Texte automatisch verarbeiten / From Form to Meaning: Processing Texts Automatically*. Narr Verlag.

Pierre Lison. 2010. Towards relational pomdps for adaptive dialogue management. In *Proceeding of the Student Research Workshop of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Pierre Lison. 2012. Probabilistic dialogue models with prior domain knowledge. In *Proceedings of the 13th SIGDIAL meeting on Discourse and Dialogue*, Seoul, South Korea. (in press).

Stéphane Ross, Joelle Pineau, Sébastien Paquet, and Brahim Chaib-draa. 2008. Online planning algorithms for pomdps. *J. Artif. Int. Res.*, 32(1):663–704, July.

S. Ross, J. Pineau, B. Chaib-draa, and P. Kreitmann. 2011. A Bayesian Approach for Learning and Planning in Partially Observable Markov Decision Processes. *Journal of Machine Learning Research*, 12:1729–1770.

D. Schlangen, T. Baumann, H. Buschmeier, O. Buß, S. Kopp, G. Skantze, and R. Yaghoubzadeh. 2010. Middleware for Incremental Processing in Conversational Agents. In *Proceedings of the 11th SIGDIAL meeting on Discourse and Dialogue*.

J. D. Williams. 2008. The best of both worlds: Unifying conventional dialog systems and POMDPs. In *International Conference on Speech and Language Processing (ICSLP 2008)*, Brisbane, Australia.

Biographical Sketch

Pierre is currently a PhD Research Fellow at the Department of Informatics of the University of Oslo (Norway). Before answering the call of the Great North and moving to Oslo, he worked for two years as a researcher at the German Research Centre for Artificial Intelligence (DFKI), where he was involved in several EU projects in cognitive robotics and human-robot interaction. He holds a M.Sc. in Computer Science & Engineering from the University of Louvain (Belgium) and a M.Sc. in Computational Linguistics from the University of Saarland (Germany).

His hobbies include reading, travelling, hiking, participating in civic/cultural projects, and salsa dancing.

Changsong Liu

Department of Computer Science and Engineering
Michigan State University
East Lansing, MI, 48864

cliu@cse.msu.edu
www.cse.msu.edu/~cliu

1 Research Interests

My research interests lie generally in the area of **human robot dialog**, with a special focus on **language grounding** and **collaborative models** in **situated dialogs**. I am also very interested in **deep language understanding** and **educational uses of dialog systems**.

1.1 Previous Work

My work in the human robot dialog area started with using the wizard-of-oz paradigm to collect data and investigate interesting phenomena. In (Liu et al., 2010), I investigated the ambiguity of spatial language in situated human robot interaction. We found that spatial language could be inherently ambiguous even in relatively simple settings. This is due to the implicit *frame-of-reference* and its interaction with situational factors such as spatial arrangements and individual preferences. My recent work investigated the impact of mismatched perceptual capabilities on collaborative dialogs and the influence of non-verbal modality, i.e. eye gaze (Liu et al., 2011; Liu et al., 2013). Using a variant version of the wizard-of-oz paradigm, we simulated mismatched visual capabilities between two human partners and collected their collaborative dialogs. We found that collaboration was significantly impacted by mismatched visual capabilities and eye gaze played an important role to facilitate interaction.

1.2 Current Work

Inspired by previous work, my current work focuses on developing computational models and algorithms to mediate between mismatched perceptual capabilities (i.e. between a human and a computer vision based agent). We currently focus on referential grounding and use inexact graph matching as the mechanism. In our method, the human speaker's referring discourses and the agent's visual perceptions are both converted to attributed relational graphs. Inexact graph matching algorithm is then applied to the two graphs to find an optimal match between the discourse referents and the visually percepted entities. Our results indicate that inexact graph matching is a promising mechanism to mediate between mismatched perceptions because of its error-tolerance nature. More details are presented in (Liu et al., 2012).

1.3 Future Work

Along with my current research focus, I will address the following research problems in my future work:

- Currently the graph matching algorithm only considers one-to-one node mapping, thus it can not handle group descriptions (e.g. plural nouns). In our data, we observed that group descriptions are common and useful. So next we will explore more advanced matching algorithms such as the hyper-graph matching algorithm to handle group descriptions.
- Language grounding models is another important aspect of our method. These models map the elements in language to the features of visual perceptions, so that the similarity between the language graph and vision graph can be evaluated. How to build grounding models for different aspects of language, such as imprecise adjectives and spatial descriptions? How to collect/find data to train the model? How to build/train a weighting scheme to combine different aspects together? These questions will be addressed in the near future.
- My current work on referential grounding is only part of our larger on-going project. Focusing on mediating between mismatched capabilities in human robot dialog, we will extend our research to language acquisition, language generation and collaborative models for dialog management. Our ultimate goal is to integrate our research advances together to build a collaborative robotic agent that engages human partners in situated dialogs.

2 Future of Spoken Dialog Research

In the next 5 to 10 years, I would expect to see more and more practical uses of dialog systems in many different areas. The commonly used "pure speech" (telephone-based) dialog systems will become more robust and intelligent. Situated dialog systems, such as speech-based assistants on smart devices, will also become very popular. Besides, I think that the advances in dialog research will soon bring substantial progress to two other areas – human robot interaction and education. Dialog systems will

provide the most natural way for human users to interact and collaborate with robots, thus social robots will begin to gain its popularity when it is equipped with advanced dialog capabilities. In the area of education, dialog-based agents that can infer the student's cognitive and emotional status will serve as efficient and friendly teachers. With their help, every student can become a more self-paced and -motivated learner.

To move more dialog systems out of research labs and into daily lives, I think we need to make continuous efforts in directions including the following:

- Deep language understanding: Understanding the speaker's intention is very important but also difficult. To achieve this, we should not only rely on one kind of methodology. Both statistical and logic-reasoning approaches should be integrated and enhance each other. How to represent common knowledge? How to automatically acquire them (e.g. from the web)? How to integrate inferences into deep semantic analysis? Those are all interesting questions.

- Cognitive systems: Beyond the telephone-based spoken dialog, now we have many other forms of dialog, such as dialog with virtual agents, dialog with personal devices and dialog with robots. Those forms of dialogs provide a great level of embodiment, thus non-verbal modalities should become an essential element. Also, as dialog systems become more "personal", we should also take factors such as emotions and long-term relations into account.

- Situated dialog. Situatedness provides both great opportunities and challenges into dialog research. It provides a very rich context for an agent to interpret language and direct the dialog. But the openness of the environment also jeopardizes "common ground" between humans and agents because of their mismatched perceptual capabilities. Thus, error-tolerate mechanisms and collaborative models that can mediate between the mismatched capabilities should play an important role in situated dialog systems.

3 Suggestions for Discussion

The topics I would like to suggest for discussion are:

- Wizard-of-oz paradigm for data collection in human robot dialogs: The advantages and disadvantages of the wizard-of-oz experiments, "ablated" wizard-of-oz design, human-in-the-loop design for dialog systems.

- Computational models for the *common ground* theory: Representation of common ground, implementation of the theory within dialog management, evaluation methods/metrics.

- The role of language acquisition: Motivations/scenarios for language acquisition, relations/differences between language acquisition and statistical learning, dialog-driven language acquisition.

References

C. Liu, J. Walker, and J.Y. Chai. 2010. Ambiguities in spatial language understanding in situated human robot dialogue. In *Proceedings of the AAAI Fall Symposium on Dialog with Robots*.

C. Liu, D.L. Kay, and J.Y. Chai. 2011. Awareness of partners eye gaze in situated referential grounding: An empirical study. In *2nd Workshop on Eye Gaze in Intelligent Human Machine Interaction*.

C. Liu, R. Fang, and J.Y. Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages –. Association for Computational Linguistics.

C. Liu, R. Fang, and J.Y. Chai. Shared gaze in situated referential grounding: An empirical study. To appear in *Eye Gaze in Intelligent Human Computer Interaction*. Springer. 2013.

Biographical Sketch

Changsong Liu is a PhD student in the Language and Interaction Research (LAIR) lab at Michigan State University. He received his B.E. and M.E. degree from University of Electronic Science & Technology of China.

He has a dream that every student in the world, especially those who live in a region that lacks good schools and teachers, can have an equal chance of being well-educated. It is his belief that the advances of dialog systems and AI technologies in general will change our education, thus change our world.

Alejandra Lorenzo

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

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

1 Research Interests

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

1.1 About Dialog Systems and SLA

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

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

My current work consists in developing such a chatbot, which is going to be integrated into a serious game for language learning developed in the context of the EU funded ALLEGRO project (`http://talc.loria.fr/-ALLEGRO-Nancy-.html`). The ALLEGRO project aims to develop web services and in particular, serious games for language learning.

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

1.2 First Experiments

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

However, the supervised approach used in the Emospeech project requires to collect and annotate new data each time we change the dialog system or the domain. I am currently investigating possible solutions to this problem. If we consider the task of semantic role labeling as (part of) the interpretation module, the goal is to develop an automatic semantic role labeler, which would serve as interpreter. So, in such a situation, two possible approaches to increment portability could be:

1. Semantic Role Projection: This would be a way to take advantage of the existence of annotations in a resource-rich language (usually English) to be projected to a resource-poor language (in this case, for SRL, French). As an example of recent work in this context we can cite (Van der Plas et al., 2011). Although they minimize the manual effort involved, these approaches still require both an annotated source corpus and an aligned target corpus. Moreover, generally they are not portable from one framework to another.

2. **Unsupervised SRL**: In this context several approaches have been proposed. (Swier et al., 2004) were the first to introduce unsupervised SRL in an

approach that used the VerbNet lexicon to guide unsupervised learning. Following this work other approaches were proposed, where the methods and algorithms vary, but the unsupervised focus not. E.g.: (Grenager et al., 2006), (Lang et al., 2010), (Lang et al., 2011a), (Lang et al., 2011b),(Titov et al., 2011), (Titov et al., 2012)

I am more interested in the second approach, and for that I started experimenting with unsupervised techniques to Semantic Role Labeling. In (Lorenzo et al., 2012) we propose an unsupervised approach to semantic role induction that uses a generative Bayesian model.

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

2 Future of Spoken Dialog Research

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

About dialog systems in general, I always thought that in order to create machine that behave like humans, we should first take a look at "what" we want to mimic and learn and behave as much as possible they way they learn. In that context, I've always felt that we do not use a single method or approach when we "learn" something new or when, during a dialog, we "understand" the information received from the other dialog participant. Instead, we use a combination of methods that we could loosely speaking classify as symbolic and statistical. In the same way, I feel that not just one, but a combination of methods are needed to overcome the problems that each one present nowadays.

3 Suggestions for Discussion

Possible topics for discussion:

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

References

Robert S. Swier and Suzanne Stevenson. 2004. *Unsupervised Semantic Role Labelling*. EMNLP 2004.

Grenager, Trond and Manning, Christopher D.. 2006. *Unsupervised discovery of a statistical verb lexicon*. EMNLP 2006.

Joel Lang and Mirella Lapata. 2010. *Unsupervised induction of semantic roles*. HLT 2010.

Joel Lang and Mirella Lapata. 2011. *Unsupervised Semantic Role Induction via Split-Merge Clustering*. ACL 2011.

Joel Lang and Mirella Lapata. 2011. *Unsupervised Semantic Role Induction with Graph Partitioning*. EMNLP 2011.

Lonneke van der Plas and Paola Merlo and James Henderson. 2011. *Scaling up Cross-Lingual Semantic Annotation Transfer*. ACL/HLT 2011.

Ivan Titov and Alexandre Klementiev. 2011. *A Bayesian Model for Unsupervised Semantic Parsing*. ACL 2011.

Ivan Titov and Alexandre Klementiev. 2012. *A Bayesian Approach to Unsupervised Semantic Role Induction*. EACL 2012.

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

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

Biographical Sketch

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

Teruhisa Misu

National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Keihanna Science City,
Kyoto, Japan

teruhisa.misu@nict.go.jp
http://mastarpj.nict.go.jp/~xtmisu/

1 Research Interests

My current research interest is in making **proactive dialog systems**. Although most current spoken dialog systems (SDSs) handle simple database retrieval or transactions driven by users' explicit requests, it is desirable that SDSs should make proactive interactions such as clarifications, recommendations and advice like a hotel concierge or an experienced operator. More specifically, I am interested in **dialog management in non-database retrieval tasks** and **adaptive response generation**.

1.1 Previous work

My previous work has focused on SDSs based on **information retrieval** and **question-answering** (QA) using **a set of documents as a knowledge base** (Misu, 2008).

1.1.1 Confirmation and Clarification in Voicesearch Tasks

It is indispensable for SDSs to interpret a user's intention robustly in the presence of speech recognition errors and extraneous expressions characteristic of spontaneous speech. In speech input, moreover, users' queries tend to be vague, and they may need to be clarified through dialog in order to extract sufficient information to get meaningful retrieval results. In conventional database query tasks, it is easy to cope with these problems by extracting and confirming keywords based on semantic slots. However, it is not straightforward to apply such a methodology to general document retrieval tasks.

To solve these problems, we proposed a confirmation method based on two statistical measures that are not based on the confidence measure of ASR, but on the impact on retrieval as well as the degree of matching with the backend knowledge base (Misu and Kawahara, 2006b; Misu and Kawahara, 2008). We also proposed a method to make clarification questions by dynamically selecting from a pool of possible candidate questions. As the criterion for the selection, the information gain is defined based on the reduction in the number of matched items (Misu and Kawahara, 2005; Misu and Kawahara, 2006b).

1.1.2 Interactive Navigation based on QA and Information Recommendation

We proposed an interactive dialog framework. In conventional audio guidance systems, such as those deployed in museums, the information flow is one-way and the content is fixed. We prepare two modes, a user-initiative retrieval/QA mode (pull-mode) and a system-initiative recommendation mode (push-mode), and switch between them according to the user's state. In the user-initiative retrieval/QA mode, the user can ask questions about specific facts in the documents in addition to general queries. In the system-initiative recommendation mode, the system actively provides the information the user would be interested in. The system utterances are generated by retrieving from and summarizing Wikipedia documents. We implemented a navigation system containing Kyoto city information. The effectiveness of the proposed techniques was confirmed through a field trial by a number of real novice users (Misu and Kawahara, 2007b; Misu and Kawahara, 2007a).

Recently, I also proposed a user model of QA dialog (Misu et al., 2012).

1.1.3 Efficient Language Model Construction for SDSs

We proposed a bootstrapping method of constructing statistical language models for new SDSs by collecting and selecting sentences from the World Wide Web. Out of the collected texts, we select "matched" sentences both in terms of the domain and in utterance style, thus appropriate for training data of the language model (Misu and Kawahara, 2006a).

1.2 Current and Future Work

Currently, we are developing consulting dialog systems that help make decisions through spontaneous interactions.

Most previous studies assumed a definite and consistent user goal. The dialog strategies were usually designed to minimize the cost of information access. However, this assumption fails in various real world situations. Specifically, we are developing a dialog system that handles tourist guidance. Thus far, we have collected itinerary planning dialogs in Japanese, in which users

plan a one-day visit to Kyoto City (Misu et al., 2009). It contains various exchanges, such as clarifications and reasonings. The user may explain his/her vague preferences by listing examples. The server would sense the users preference from his/her utterances and then request a decision.

In order to construct a consulting dialog system from the corpus, we are annotating dialog acts (Misu et al., 2009), then proposed a dialog state model in such consulting dialog (Misu et al., 2010).

My recent work also includes a spoken dialog interface that evokes spontaneous user reactions (Misu et al., 2011).

2 Future of Spoken Dialog Research

I am afraid that the more multifunctional cell phones become (e.g. iPhone, Android), the less users use speech interfaces for simple query tasks such as train search, hotel reservation, etc. Thus we will have to propose applications that can make use of something that such smartphones do not have. It may be a large (human-sized) touchscreen or a motion detection sensor. Another direction would be the expansion of tasks that a speech interface can handle (but a small touch screen cannot).

3 Suggestions for discussion

- **How to evaluate dialog systems**
  What is a good evaluation measure for dialog systems? Can we define a evaluation measure like "BLEU/NIST score" for machine translation. (especially in non-goal-oriented dialog systems)
- **How to make users behave naturally as a human operator can.**
  Users talk to the systems in different utterance style from the style they talk to human operators. What are the causes of this phenomenon?
- **Cost reduction in developing new SDSs using the WWW**
  How can we make use of web resources for the construction of SDSs.

References

T. Misu and T. Kawahara. 2005. Speech-based information retrieval system with clarification dialogue strategy. In *Proc. Human Language Technology Conf. (HLT/EMNLP)*.

T. Misu and T. Kawahara. 2006a. A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts. In *Proc. Interspeech*, pages 9–12.

T. Misu and T. Kawahara. 2006b. Dialogue Strategy to Clarify User's Queries for Document Retrieval System with Speech Interface. *Speech Communication*, 48(9):1137–1150.

T. Misu and T. Kawahara. 2007a. An interactive framework for document retrieval and presentation with question-answering function in restricted domain. In *Proc. Int'l Conf. Industrial, Engineering & Other Applications of Artificial Intelligent Systems (IEA/AIE)*, pages 126–134.

T. Misu and T. Kawahara. 2007b. Speech-based Interactive Information Guidance System using Question-Answering Technique. In *Proc. ICASSP*.

T. Misu and T. Kawahara. 2008. Bayes risk-based dialogue management for document retrieval system with speech interface. In *Proc. COLING, Vol.Posters & Demo*, pages 59–62.

T. Misu, K. Ohtake, C. Hori, H. Kashioka, and S. Nakamura. 2009. Annotating Communicative Function and Semantic Content in Dialogue Act for Construction of Consulting Dialogue Systems. In *Proc. Interspeech*.

T. Misu, K. Sugiura, K. Ohtake, C. Hori, H. Kashioka, H. Kawai, and S. Nakamura. 2010. Dialogue Strategy Optimization to Assist User's Decision for Spoken Consulting Dialogue Systems. In *Proc. IEEE-SLT*, pages 342–347.

T. Misu, E. Mizukami, Y. Shiga, S. Kawamoto, H. Kawai, and S. Nakamura. 2011. Toward Construction of Spoken Dialogue System that Evokes Users' Spontaneous Backchannels. In *Proc. SIGDIAL*.

T. Misu, K. Georgila, A. Leuski, and D. Traum. 2012. Reinforcement Learning of Question-Answering Dialogue Policies for Virtual Museum Guides. In *Proc. SIGDIAL*.

T. Misu. 2008. *Speech-based Navigation Systems based on Information Retrieval and Question-Answering with Optimal Dialogue Strategies*. Ph.D. thesis, School of Informatics, Kyoto University.

Biographical Sketch

Teruhisa Misu is a researcher in the Spoken Language Communication Laboratory, at National Institute of Information and Communications Technology (NICT), Japan. He received the B.E. degree in 2003, the M.E. degree in 2005, and the Ph.D. degree in 2008, all in information science, from Kyoto University, Kyoto, Japan. From 2005 to 2008, he was a Research Fellow (DC1) of the Japan Society for the Promotion of Science (JSPS). In 2008, he joined NICT Spoken Language Communication Group. From 2011/11 to 2012/2, he was a Visiting researcher at USC/ICT.

Christopher M. Mitchell

Department of Computer Science
North Carolina State University
890 Oval Drive
Raleigh, NC 27695

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

1 Research Interests

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

1.1 Studying Effective Tutorial Dialogue

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

1.2 Convergence and User Adaptation

Convergence, the phenomenon of humans becoming more similar in their lexical, prosodic, and multimodal behaviors over time, has been widely studied, both within the domain of tutorial dialogue and in other domains. While a link between convergence and task success has been suggested in non-tutorial domains (Stoyanchev and Stent 2009), the results are not as clear for tutorial dialogue. For example, Ward and Litman (2007) found that lexical convergence was positively associated with learning gains for students with low pretests in a corpus of human-human tutoring, while Steinhauser et al. (2011) found that tutor-mimicking was negatively associated with learning in a corpus of human-computer tutoring. Thus, to better understand the role of convergence within tutoring, I have examined lexical convergence in the JavaTutor corpus (Mitchell et al., 2012b). The results indicate a longitudinal trend: users were more likely to reuse their partners' words as they engaged in more dialogues together, with a significant increase observed between the first and sixth tutoring session. Several measures of convergence were also predictive of specific aspects of both dialogue success and user affect. For example, students who reused tutor words at higher rates reported that the tasks were less mentally demanding, and tutors who reused student words at higher rates were found to be less effective at producing learning gains. These results highlight the potential for applying convergence analysis to create more effective tutorial dialogue system adaptation.

1.3 Future Directions

Building on my work with learning effective tutorial dialogue strategies from data, I plan to further analyze these strategies by using fine-grained task success rather than overall learning gains to measure effectiveness of specific sub-dialogues. I also plan to expand the analysis to the entire corpus, which requires the development of automatic dialogue act annotation approaches. The ultimate goal of this dialogue modeling work is to learn effective dialogue management models. To accomplish this goal I plan to utilize machine learning frameworks such as hierarchical hidden Markov models and reinforcement learning. Automatically learning effective strategies at runtime, to create a real-time adaptive dialogue policy, is also an important direction for my future research.

I plan to build on my work on convergence in dialogue by investigating the phenomenon along other dimensions; for example, whether the apparent affective state of one user influences the apparent (or actual) affective state of the other user. I am also interested in investigating whether these convergence phenomena transfer to human-computer interaction; that is, whether a user is more or less likely to converge to an automated dialogue partner than to a human dialogue partner.

Finally, I am interested in the complexities of mixed initiative for tutorial dialogue systems; specifically, looking not just at what actions a dialogue system should provide, but when to undertake these actions and when to let the user lead the interaction instead. This issue is particularly important in tutoring, as a poorly timed intervention may decrease the efficiency of learning.

2 Future of Spoken Dialog Research

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

3 Suggestions for Discussion

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

References

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

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

Natalie B. Steinhauser, Gwendolyn E. Campbell, Leanne S. Taylor, Simon Caine, Charlie Scott, Myroslava O. Dzikovska, and Johanna D. Moore (2011). Talk Like an Electrician: Student Dialogue Mimicking Behavior in an Intelligent Tutoring System. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 361-368.

Svetlana Stoyanchev and Amanda Stent (2009). Lexical and Syntactic Priming and Their Impact in Deployed Spoken Dialog Systems. In *Proceedings of NAACL HLT*, 189-192.

Arthur Ward and Diane Litman (2007). Dialog Convergence and Learning. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 262-269.

Biographical Sketch

Christopher M. Mitchell is a Ph.D. student at North Carolina State University, where he is advised by Kristy Elizabeth Boyer and James C. Lester. Chris holds a B.S. in Computer Science with minors in Cognitive Science and Mathematics from North Carolina State University. His research interests lie in Computational Linguistics, Dialogue Management, and Intelligent Tutoring Systems.

Hyungjong Noh

Pohang University of Science & Technology
Dept. of Computer Science & Engineering
790-784 San 31 Hyoja-dong Nam-gu Pohang
Gyungbuk South Korea

nohhj@postech.ac.kr

1 Research Interests

My research interest is beyond task-oriented dialog. In other words, I am interested in how to process more **colloquial dialogs** including **chat-like conversation**. Previous researches generally focused on maintaining robustness of the spoken dialog system (SDS) and high task completion rate. However, as well as task-oriented dialogs, the needs of processing other free conversations also are increasing. For example, we can build an SDS for **language learning** purpose. In this case, diversity of utterance expression or intentions is more important than consistency of system response. Even if the dialog state is the same, the system would have to different responses, so that the user (learner) can be familiar to various expressions. Of course, task-oriented corpus can be used for language learning SDS to train the learner for specific situations such as ticketing, reserving, etc. However, we should concentrate on different points from traditional SDS as described above. Building the chatter bot is another problem. Generally, chatting conversation has no specific purposes. A user just speaks to the system for entertainment or fun. Many previous SDSs cannot treat this because they assume that dialogs have certain purposes and they try to help users to fulfill the purposes. Therefore, I have been developing new dialog systems for more natural and flexible dialogs using ranking several intention-related features and **user help guide and suggestion**. Through the systems, we can conduct dialogs which have relatively many intentions and their n-gram combinations, and give users feedback which can be used for language learning as well as task completion.

1.1 Intention-related features

For the past several years, I have developed SDSs which can treat chat-like dialogs. I consider dialogs which have many intention types, their combinations as chat-like dialogs. The intentions contain many trivial expressions, such as “great”, “let me see”, unrelated to specific tasks. I have focused on how to process these various intentions and have suggested some policies which can calculate discourse similarity effectively. I started developing from instance-based SDSs (Lee et al., 2008) first, and adopted the discourse similarity feature to the system (Noh et al., 2011b).

In addition to the similarity feature, I defined other features about causal relation between intentions and entity slot filling state. To aggregate these, I used ranking SVM algorithm. Therefore the current system can be viewed as a hybrid system, which is an instance-based SDS and adopted statistical methodology.

1.2 User help guide and feedback

For novice users, using SDSs can be somewhat difficult. User utterances could be misrecognized due to the current limitation of ASR technology, or users would give improper utterance to the current dialog state. In these situations, the dialog flow could be misled without some help guide or feedback from the systems. When considering language learning SDSs, learners would need some suggestions about current possible expressions or some hints to make their own utterances. In other words, appropriate feedback can help users for various purposes. Therefore I am developing some feedback mechanisms that give useful information to users. For each turn in dialogs, user utterance suggestion could help users to make proper utterances to the current dialog state. This function is used for both task-oriented and language learning dialogs. The system also gives feedback for users’ improper utterances according to causal relation. For example, when a user asks for the distance without destination specified in path finding domain, the system could give feedback which implies that the user should specify the destination first before asking the distance. Users can recognize their mistakes correctly from this logical feedback, and then it helps dialogs to be completed without problems.

1.3 Field Study

To investigate the effects of our approaches, especially on language learning, we performed a field study at an elementary school. Our developed SDS has been integrated into 3D environment educational game (Noh et

al., 2011a), and that game is distributed to students. We noticed that they are very interested in learning course with the dialog systems. Also the feedback function was helpful for making proper expressions to the dialog state. From the experiments, we learned that our SDS which conducts flexible dialogs and gives feedback can perform a role for language learning.

2 Future of Spoken Dialog Research

Many SDS have been developed for information-seeking purpose. However, I think that SDSs can be used for other various purposes, including language learning, colloquial conversations for sentimental support or fun.

Especially, emotional and sentimental aspects should be focused as future research topic. To do this, we may consider other measures rather than traditional task completion rate or average turn length. I am concentrating on issues on language learning and chatter bot now, but other interesting topics can be found as new SDS domain.

3 Suggestions for discussion

- Gathering 'real' chatting corpus effectively with avoiding privacy issue and crawling labor
- User modeling on free conversations rather than information-seeking conversations
- Language learning application with SDS technologies
- Searching fields to which SDS technologies can contribute and create a synergy effect

References

Cheongjae Lee, Sangkeun Jung, Gary Geunbae Lee. 2008. Robust dialog management with n-best hypotheses using dialog example and agenda, Proceedings of the 45th annual meeting of the association for computational linguistics: human language technologies (ACL:HLT), Ohio

Hyungjong Noh, Kyusong Lee, Sungjin Lee, Gary Geunbae Lee. 2011. POMY: a conversational virtual environment for language learning in POSTECH, 12th sigdial workshop on discourse and dialog, Portland (demo presentation)

Hyungjong Noh, Sungjin Lee, Kyusong Lee, Gary Geunbae Lee. 2011. Ranking dialog acts using discourse coherence indicator for English tutoring dialog systems, Proceedings of the 3rd international workshop on spoken dialog systems technology (IWSDS 2011), Granada

Biographical Sketch

Hyungjong Noh received the B.S. degree in Computer Science Engineering from Pohang University of Science and Technology in 2005. He is currently a Ph.D. student of the same university. His research interests include dialog management, intelligent computer-assisted language learning, and general spoken dialog system.

Elnaz Nouri

Institute for Creative Technolgies
12015 Waterfront Dr
Playa Vista, CA 90094

[nouri@ict.usc.edu]
http://www-scf.usc.edu/~enouri/

1 Research Interests

My current research focus is on dialogue decision making models for negotiating agents and persuasive arguments.

1.1 Summary of Past Research Work

I have worked on augmenting the performance of question answering conversational characters by using sets of linked question-answer pairs which are generated automatically from texts on different topics.

I have also developed a computational model that simulates decision making by taking into account culture. The simulations and implementation of the model in to our virtual agents show that we are able to make a more realistic model human behavior by using such model. (Our virtual agents are animated computer generated intelligent characters capable of interacting and chatting with human users). This model offers benefits for the understanding the dynamics of utility based decision making and the effect of culture on the process.

1.2 Summary of Current Research Work

I have been working on learning culture-specific weights for a multi-attribute model of decision-making in negotiation, using Inverse Reinforcement Learning (IRL). The model takes into account multiple individual and social factors for evaluating the available choices in a decision set, and attempts to account for observed behavior differences across cultures by the different weights that members of those cultures place on each factor.

2 Future of Spoken Dialog Research

I Few basic questions brought up the decision making model which I believe sits at the heart of a negotiating agent and I would be addressing them in my future work. A few examples of the questions still roaming in my head are: 1) What makes a change in your evaluation of the same item after you have been persuaded by the other person. 2) What are the differences in the strategies and language that people use when they negotiate? 3) How does our focus of attention changed when we calculate the utility? 4) Why aren't these paradigms working the same for all the people?

3 Suggestions for discussion

Here are some of suggestion for the discussion topics for the workshops:

- Future of application of the statistical methods, what are the limits and challenges that we face today
- What are the benefits and challenges of the Question-Answering systems
- Are there new domains for applying Spoken Dialogue Systems?

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. In Proceedings of the 21st International Conference on Machine Learning (ICML).Banff, Canada.

Buchan, N. R., Croson, R. T. A., & Johnson, E. J. (1999). Understanding what's fair: Contrasting perceptions of fairness in ultimatum bargaining in Japan and the United States. In Discussion paper, University of Wisconsin.

Camerer, C. F. (2003). Behavioral game theory – Experiments in strategic interaction. Princeton University Press.

Gal, Y., Pfeffer, A., Marzo, F., & Grosz, B. J. (2004). Learning social preferences in games. In Proceedings of the 19th National Conference on Artificial Intelligence (p. 226-231). San Jose, CA.

Georgila, K., Henderson, J., & Lemon, O. (2006). User simulation for spoken dialogue systems: Learning and evaluation. In Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP). Pittsburgh, PA.

Biographical Sketch


Elnaz Nouri is a PhD student at the University of Southern California computer science department. She has been a member of Natural Dialogue group at Institute for Creative Technologies since 2011 and works under the supervision of Professor David Traum.

Aasish Pappu

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
aasish@cs.cmu.edu
www.cs.cmu.edu/~apappu

1 Research Interests

I work on **Knowledge Acquisition through Spoken Dialog**. My focus is on robots that acquire new knowledge from humans about their environment through speech. Teaching robots is an important problem because it may not be feasible for a normal user to manually program a robot to perform new tasks in a new environment. Our goal is to develop an interactive learning mechanism that lets a user instruct new tasks to the robot in spoken language. For e.g., `go and clean under the sofa`. We look at this problem in navigation domain where a user can orally teach new path plans and give information about unknown objects to visually-blind robots.

The process of knowledge acquisition involves functions like instruction understanding, knowledge extraction, knowledge representation and feedback. Some of these functions are already implemented in our knowledge acquisition framework for dialog systems. This framework is currently used in Teamtalk navigation dialog system. I will discuss some of my previous work and current work in the following sections.

1.1 Understanding and Knowledge Representation of Directions

We want the mobile robots to understand natural language instructions to help people in variety of scenarios such as evacuation, domestic-help and escort robots. In order to understand how people give route instructions to each other we conducted an experiment where we asked people to give route instructions from one place to another. We have looked at how people give instructions and modify them when faced with obstacles. We observed that people give and modify instructions with no structural differences in the discourse. However they insert cautionary statements to reinforce the status of the route (e.g., `you are standing in front of an obstacle`).

From our analysis of corpus, we designed a taxonomy of route instructions. The taxonomy contains four higher level categories viz., Imperatives, Advisory, Meta Comments, Grounding statements and 18 sub-categories.

We used this taxonomy as a design rubrick for a CFG grammar for language understanding and an OWL ontology for knowledge representation for a robot. The CFG grammar largely contains concepts related to different types of locations, e.g., `hallways, floor-transitions, landmarks`. Whereas, the ontology contains representation of the robot's environment in descriptive logic. We will focus on ontology and its interaction with the dialog system in the following subsection.

1.2 Preliminary work on Knowledge Acquisition

A robot's environment can be understood as a two layered map: physical map and semantic map. Physical map is an occupancy grid with rudimentary information about walls and no-walls determined by the robot's sensors. Whereas a semantic map is a structured representation of the environment serves as a intermediate medium between natural language instructions and low-level actions of a robot. This representation acts like a fore-brain of the robot in the following ways:

1)Episodic memory: time-aligned sequence of robotic actions

2)Procedural memory: stores route plans from $A \rightarrow B$

3)Semantic memory: whereabouts of locations and status of actions

Predicates: Relationship between concepts e.g., part-whole relation

The ontology component serves as status of the world for a robot. It interacts with the human through the dialog system. The interaction involves editing procedural memory, recording new events to episodic memory, or updating the location coordinates in the semantic memory. Besides recording plans, procedural memory plays major role in scheduling and the execution of actions. It interacts with the robot through dialog interface as if a human is giving an instruction. This approach allows the human to negotiate any inconsistencies or changes in the recorded plan through dialog, unlike if a robot directly accesses the procedural memory.

1.3 Current Work: Towards Better Representation of Knowledge

Our aim is to interactively learn a semantic map of the robot's environment. This map would serve as a talking point between the human and the robot. This graph

would contain locations or landmarks as nodes and hallways, bridges as edges.

1.3.1 Semantic Map

A semantic map serves as a representation of the environment and helps us establish physical context for a natural language instruction. A semantic map can help quantify salience of landmarks mentioned in the discourse based on their proximity. A semantic map would not only contain physical attributes but also contains co-occurrence probabilities of entities. This map can be built using knowledge acquisition strategies discussed in the next subsection.

1.3.2 Acquisition Strategies

In order to acquire right information, a robot should present a right context with right level of detail. A robot cannot be too detail or too terse while asking for knowledge. Therefore, we would like the robot to use following strategies to acquire knowledge:

1) Primary: Given a route instruction, find an unidentifiable concept, then ask attributes of that concept. For e.g., distance and orientation of a location A from current position of the robot.

2) Auxillary: First follow primary strategy, then ask relation of the concept with other concepts in the physical proximity. Since, we are interested in navigation domain, physical proximity is an important criterion to acquire additional knowledge. For e.g., is this location connected to another location via hallway or How close are the locations A and B?

3) Implicit: If a route instruction informs that a particular location A is part of another location X then infer that all routes leading X will also lead to A. This can help us infer the part-whole relationships in underspecific route instructions.

1.3.3 Evaluation

Like any system we are interested in quantifying the overall performance of a knowledge acquisition system. This evaluation encompasses various aspects of the system, including speech recognition, parsing of recognized output, detecting unknown concepts, acquiring/updating knowledge, and finally execution of an instruction.

There are already evaluation metrics for some of these aspects viz., speech recognition, parsing and execution of instruction. However, detecting unknown concepts, seeking for new knowledge in a right amount of detail is Subjective. At a snapshot of knowledge, ideally, the robot should seek same level of detailed information as a human would do. We plan to understand how humans would seek new information to progress towards their end goal, whether it is navigation or some other task.

1.4 Applications of Knowledge Acquisition

Knowledge acquisition through spoken dialog can be of great utility in the following applications.

1.4.1 Recommended routes

Sometimes people prefer one route to another when there exists multiple routes between two places. Some routes would have lesser traffic, fewer stairs etc., that makes them preferrable to other routes. This kind of information is not readily available in automated path planning mechanisms, therefore this information can be augmented information through dialog.

1.4.2 Referential Ambiguity

Another relevant application is in resolving ambiguity. In navigation domain, ambiguity mainly arises due to two reasons: 1) multiple-references same entity, 2) same reference multiple entities. In the former case, the ambiguity happens to be at the lexical level, which can be resolved via left-right context in an utterance. However in the latter case, it needs to be resolved based on physical context. The semantic memory of the robot should help us resolve this ambiguity by constructing appropriate questions to the user.

2 Future of Spoken Dialog Research

In another 5-10 years SDS research should be able handle following issues:

- Multi-lingual spoken dialog systems in countries like India.
- Detect and recover from out-of-domain words and phrases.
- How to better communicate limitations and capabilities to a new user?

3 Suggestions for Discussion

- Self-schooled dialog systems: Making new connections from unexpected (yet frequent) inputs patterns.
- Subjective and Objective evaluation: Rapid system building and evaluating dialog systems at large scale.

Biographical Sketch

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

Ethan O. Selfridge

Center for Spoken Language Understanding
Oregon Health & Science University
20000 NW Walker RD, Portland, OR, USA

selfridg@ohsu.edu
http://www.csee.ogi.edu/~selfridg/

1 Research Interests

My research surrounds the improvement of turn-taking for spoken dialogue systems. This has led to three distinct research areas: importance-driven turn-bidding, incremental speech recognition, and temporal simulation.

1.1 Importance-Driven Turn-Bidding

Importance-driven turn-bidding (IDTB) refers to the system basing its turn-taking decisions on the importance of the utterance. This is motivated by psycholinguistic research that suggests the importance of the utterance is a primary factor to determining the turn-taking behavior of the potential speaker (Yang and Heeman, 2010). Other studies have shown that the turn-taking process is highly negotiative and that conversants work with each other, using a number of turn-taking cues, to determine who has the floor. In 2010 I proposed the IDTB framework where, using simulation, both conversants used turn-bids (modeled as utterance onset) to compete for the turn (Selfridge and Heeman, 2010). The user employs a stochastic process to select the turn-bid, whereas the System used reinforcement learning to select both the utterance and the turn-bid; indirectly learning utterance importance. I found that an IDTB turn-taking framework was far more flexible and efficient than more conventional methods, suggesting that future systems should be capable of this type of behavior.

1.2 Incremental Speech Recognition

In order to make turn-taking decisions, the System *must* have a way to know what the user is saying as they are saying it. With this situational knowledge, the System can judge whether to interrupt, start speaking just as the user finishes, or even wait to speak in anticipation of a user's repetition. Incremental Speech Recognition (ISR), where recognition results are returned prior to complete decoding, has been used to access user speech as they are speaking.

However, ISR is not without difficulty. One central challenges to ISR is stability. Stability refers to the propensity of partials to change as decoding progresses. One can increase stability by delaying recognition but this increases the recognition lag, thus decreasing the efficacy of ISR. In 2011 I presented a paper that tackled the challenge of partial stability (Selfridge et al., 2011). The paper first presents Lattice-Aware ISR, which uses the lattice-structure to determine when to return a partial or not. This increases the stability of partials by combining low-occurring but completely stable Immortal partials with high-occurring but highly unstable Terminal partials. It then proposes the use of logistic regression to predict the stability of partials, and demonstrates that this method has significantly more discriminative ability than generic confidence scores (for both stability and accuracy). At SIGdial 2012 I am presenting a paper that first proposes an Incremental Interaction Manager that enables non-incremental SDS to garner some of the benefits of ISR, and then shows that the IIM can be used to integrate ISR with a POMDP dialogue system.

1.3 Temporal Simulation

In order to train a dialogue system with any degree of sophistication, simulations must be used. Current turn-taking simulations are either completely focused on prosody or are too stylized for training deployable systems. At SIGdial 2012 I am presenting a first attempt Temporal Simulator for SDS. This simulator models both the timing and content of user and system speech, as well as the ISR/VAD components that are the inputs to the dialogue manager. The simulator is demonstrated by comparing three different turn-taking strategies, one that is conservative and never interrupts, one that is aggressive and will interrupt frequently, and one that uses ISR confidence scores (i.e. dialogue context) to choose between being conservative or aggressive turn-taking. The simulator shows the context-based approach can maintain efficient interactions under conditions of poor ASR performance, but minimize interruptions when ASR performance increases. One of the most difficult components to simulate is incremental speech recognition, which must be simulated so that acoustic features and confidence scores are characteristic of authentic ISR. I am currently working on synthesizing ISR results, and I hope to share some very preliminary but exciting results in conjunction with the poster.

2 Future of Spoken Dialog Research

- Where do you think the field of dialogue research will be in 5 to 10 years?
  Dialogue research will be interested in complex reasoning, and will be implementing many concepts that have only been theoretical. I believe that emergent human behavior and human mimicry will be hot areas and dialogue research will also be concerned with developing "persistent" systems that learn and remember over a long period of time. Models of memory, behavior, and knowledge adaptation will be better developed and the field itself will be even more interdisciplinary than it already is.

- What do you think this generation of young researchers could accomplish in that time?
  I think that this is *the* generation of researchers that will push dialogue over the top. Never before have we had so many resources and so many good people. Speech recognition, the consistent limiting factor, is finally somewhat usable and theoretical concepts are finally being able to be tested due to computational availability. Dialogue researchers will produce systems that can back-channel at the appropriate moment, adapt operating behavior to the user and the environment, and be able to act confidently using poor ASR. They will produce systems that will "live" in the cloud and be interacting with thousands of users a day. This will give rise to new methods of harnessing and exploiting user data, something that is relatively lacking today.

- What kind of questions need to be investigated to get the field to that point?
  I think the field as a whole needs to focus on developing dialogue systems that are robust to ASR errors. Since ASR is very usable now, we need to have systems that can cope with the errors that do occur such as during background noise. We also need to be working on more flexible systems that leverage expert knowledge and machine learning to deliver the best possible user experience.

3 Suggestions for Discussion

- Effective re-branding of SDS: Discussing ideas to change the stigma against dialogue systems
- Crowd-sourcing SDS development: Is it possible and how would it work?
- Autonomous Starbucks: domains that are low-hanging fruit for SDS but not yet used

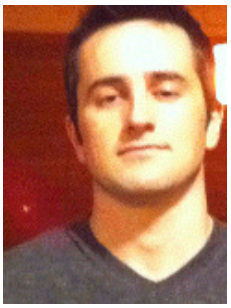
References

E.O. Selfridge, I. Arizmendi, P.A. Heeman, and J.D. Williams. 2011. Stability and accuracy in incremental speech recognition. In *Proceedings of the SIGdial 2011*.

E.O. Selfridge and P.A. Heeman. 2010. Importance-Driven Turn-Bidding for spoken dialogue systems. In *Proc. of ACL*, pages 177–185.

Fan Yang and Peter A. Heeman. 2010. Initiative conflicts in task-oriented dialogue". *Computer Speech Language*, 24(2):175 – 189.

Biographical Sketch



Ethan received a BA in Psychology from Reed College in 2006, where he completed an undergraduate thesis on cognitive dissonance and memory errors. He then worked for a speech start-up from late 2006 till the Fall of 2008 developing dialogue systems. Since the Fall of 2008 he has been working towards a PhD in Computer Science and Engineering at the Center for Spoken Language Understanding at Oregon Health & Science University. His advisor is Dr. Peter Heeman.

Seonghan Ryu

Pohang University of Science and Technology
Dept. of Computer Science and Engineering
790-784 San 31 Hyoja-dong Nam-gu Pohang
Gyungbuk South Korea

ryush@postech.ac.kr
www.about.me/ryush

1 Research Interests

My research interests lie in multi-domain spoken dialog system (SDS). Specifically, I am interested in **domain selection**, **discourse management**, and **cooperative architecture** of multi-domain SDS. Currently, the goal of my main project is to implement a multi-domain SDS which helps user with searching various contents and controlling the device.

1.1 Domain Selection

Domain selection is a bottleneck of the performance of multi-domain SDS because domain specific natural language understanding (NLU), dialog management (DM), and natural language generation (NLG) are executed after a domain is selected. Current goal in the project is to detect multiple domains for a single request in order to provide multi-domain services simultaneously.

1.2 Discourse Management

When multiple domains are selected and corresponding domain experts share some common slots, discourse becomes more complex and can no longer be managed by simple heuristic rules. Therefore, we need a new framework and strategies for multi-domain discourse management.

1.3 Cooperative Architecture

Traditional multi-domain dialog systems have adopted distributed architecture for high domain extensibility (Lin et al., 1999). But in some cases, cooperation with domain experts in NLU, DM, and NLG can generate better response for a single request. Furthermore, cooperative architecture is a solution for simultaneous multi-domain SDSs.

2 Future of Spoken Dialog Research

Since a dialog is the most natural method of communication for human, I am expecting to voice user interface to open a new prospect in the field of human computer interaction. However, it seems that it is impossible for dialog to be the major user interface in a decade. The most challenging issues of spoken dialog research are recognition and understanding natural language. Even state-of-the-art automatic speech recognition (ASR) and NLU cannot completely recognize and understand user utterances because of environmental noise, personal variances, grammatical errors, etc. I think recognition and understanding performance should reach the critical point for SDS to be a major user interface.

Robust dialog management is another challenging problem. In relatively easier task such as contents search, SDS can control dialog flow using heuristic rules. However, we need more advanced dialog modeling techniques for complex task and flexible dialog flow. Data-driven dialog modeling techniques are necessary (Lee et al., 2009) because it is unrealistic to build handcrafted rules for dialog flow. Furthermore, SDS should be able to deal with uncertainty (Williams et al., 2007; Kim et al., 2010). Capability of handling uncertainty is expected to overcome recognition or understanding errors mentioned above. In addition, we need an additional method for reducing the complexity (Kim et al., 2008) to overcome complexity problem.

3 Suggestions for discussion

- *Domain knowledge*: there is more to SDS than linguistic knowledge. How do we represent domain knowledge in a formal way and share the knowledge with the outside world?
- *Overnight training*: commercial services acquire tons of log data from real users every day. How do we use them effectively for SDS to improve every day without human efforts?
- *Out-of-domain requests*: chatting is the most natural way to handle out-of-domain requests. How do we manage dialog flow naturally between task-oriented dialog and chat dialog?
- *Usability for developers*: how do we provide more general and easy-to-build interfaces to SDSs for developers?
- *Computer game*: playing computer game with voice will be exciting experiences. What will

be the next generation computer game using the voice user interface?

- *Commercials*: currently the most successful and famous dialog system software built into the most successful and famous smart phone has been sued for its "deceptive" commercials. How do SDSs attract public gaze without being embroiled in a legal battle?

References

Bor-shen Lin, Hsin-min Wang, and Lin-Shan Lee. 1999. *A Distributed Architecture For Cooperative Spoken Dialogue Agents With Coherent Dialogue State And History*. Proceeding of IEEE Workshop on Automatic Speech Recognition and Understanding.

Cheongjae Lee, Sangkeun Jung, Seokhwan Kim, and Gary Geunbae Lee. 2009. *Example-based Dialog Modeling for Practical Multi-Domain Dialog System*. Speech Communications, 51:5 (466-484).

Jason D. Williams and Steve Young. 2007. *Partially Observable Markov Decision Processes for Spoken Dialog Systems*. Computer Speech and Language, 21:2 (393-422).

Kyungduk Kim, Cheongjae Lee, Sangkeun Jung, and Gary Geunbae Lee. 2008. *A Frame-based Probabilistic Framework for Spoken Dialog Management using Dialog Examples*. Proceeding of the 9th Sigdial Workshop on Discourse and Dialog.

Kyungduk Kim, Cheongjae Lee, Donghyun Lee, Junhwi Choi, Sangkeun Jung and Gary Geunbae Lee. 2010. *Modeling Confirmations for Example-based Dialog Management*. Proceedings of the 2010 IEEE Workshop on Spoken Language Technology.

Biographical Sketch

Seonghan Ryu received the B.S. degree in Computer Science & Engineering from Dongguk University in 2012. He is currently a Ph.D. student of Pohang University of Science and Engineering under the supervision of Prof. Gary Geunbae Lee. His research interests include multi-domain spoken dialog system, knowledge engineering, and machine learning. He has no particular hobby to mention.

William Yang Wang

School of Computer Science
Carnegie Mellon University, USA

ww@cmu.edu
www.cs.cmu.edu/~yww

1 Research Interests

I have broad interests in **statistical machine learning**, **computational linguistics**, and **spoken dialogue systems**. I am currently interested in the following research areas:

- Rapport building for life-long spoken dialogue systems.
- Topic modeling, mixed-effects latent variable models, and sparsity.
- Spoken language understanding, paralinguistics, and speech synthesis.

1.1 Using Sparse Log-Linear Models to Build Positive (and Impolite) Relationships with Teens

Spoken dialogue systems that are built for long-term interaction with one user must know how to adapt their language as the system becomes more familiar over time. Part of this challenge involves building and signaling aspects of long-term relationships, such as rapport, and using the contextually appropriate linguistic devices to do so. For tutorial systems, this challenge may additionally require knowing how rapport-building proceeds among non-adult users. In a recent paper (Wang et al., 2012), we therefore investigate the conversational strategies used by teenagers in peer tutoring dialogues, and their effects on a friend or stranger partner. In particular, we use annotated and automatically extracted linguistic devices to predict impoliteness and positivity in the next turn, using Lasso, ridge estimator, and elastic net based composite penalty log-linear models. We evaluate the predictive power of our models under various settings, and compare our sparse models with standard non-sparse solutions. Our experiments demonstrate that our models are more accurate than non-sparse models quantitatively, that tutors and tutees, and friends and strangers, demonstrate quite different patterns of how talk fulfils positive and negative social functions.

1.2 Sparse Mixed-Effects Latent Topic Models

Discovering topical information in spoken dialogues is of paramount significance for both human-human dialogue understanding, as well as spoken dialogue systems. We propose a latent variable model to enhance topic modeling. This work extends prior work in topic modelling by incorporating metadata, and the interactions between the components in metadata, in a general way. To test this, in a recent study (Wang et al., 2012b), we collect a corpus of slavery-related United States property law judgements sampled from the years 1730 to 1866. We study the language use in these legal cases, with a special focus on shifts in opinions on controversial topics across different regions. Because this is a longitudinal data set, we are also interested in understanding how these opinions change over the course of decades. We show that the joint learning scheme of our sparse mixed-effects model improves on other state-of-the-art generative and discriminative models on the region and time period identification tasks. Experiments show that our sparse mixed-effects model is more accurate quantitatively and qualitatively interesting, and that these improvements are robust across different parameter settings. Our model is also applicable to the domain of spoken dialogue understanding and mining topics in spoken dialogue systems.

1.3 Spoken Language Understanding, Paralinguistics, and Speech Synthesis

Since automatic speech recognition techniques are still far from perfect, a challenging issue for almost all spoken dialogue system is to build reliable and robust spoken language understanding components. Together with USC researchers (Wang et al., 2011), we have investigated the phonetic and lexical mixture models for spoken language understanding.

During my masters study at Columbia, I have also actively involved in modeling speaker states and paralinguistics for spoken dialogue systems. We have studied various lexical, prosodic, and phonetic approaches for intoxication detection (Wang et al., 2012c). We have also investigated multistream prediction feedback based fusion approaches for modeling level-of-interest of speakers (Wang and Hirschberg, 2011).

Besides the understanding components, I have also worked on problems related to speech synthesis for SDS. As we know, unit-selection is one of the two dominating approaches for speech synthesis nowadays. Comparing to HMM based synthesis, unit-selection approaches might be more natural, but it requires large speech corpus

to improve the coverage and generalization of the voices. To relax this problem, we have studied automatic methods to understand word level synthesis errors (Wang and Georgila, 2011).

2 Future of Spoken Dialog Research

The future generation of spoken dialogue research must solve the problem of inference under uncertainty, and must know how to make friends with users in a long term.

- Life-long spoken dialogue systems.

3 Suggestions for Discussion

I would like to see the following discussions in this year's YRRSDS:

- Latent variable models and structured sparsity for spoken dialogue systems.
- Emotions and user modeling in spoken dialogue systems.

References

William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black and Justine Cassell. 2012. "Love ya, jerkface": using Sparse Log-Linear Models to Build Positive (and Impolite) Relationships with Teens. *To appear in Proceedings of the 13th annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2012)*. Seoul, Korea, July 5-6, ACL.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012b. Historical Analysis of Legal Opinions with a Sparse Mixed-Effects Latent Variable Model. *To appear in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*. Jeju Island, Korea, July 8-14, ACL.

William Yang Wang, Fadi Biadsy, Andrew Rosenberg, and Julia Hirschberg. 2012c. Automatic Detection of Speaker State: Lexical, Prosodic, and Phonetic Approaches to Level-of-Interest and Intoxication Classification. *in Computer Speech and Language*. Elsevier.

William Yang Wang and Kallirroi Georgila. 2011. Automatic Detection of Unnatural Word-Level Segments in Unit-Selection Speech Synthesis. *in Proceedings of the ASRU 2011*. Big Island, Hawaii, USA, Dec. 11-15, 2011.

William Yang Wang, Ron Artstein, Anton Leuski, and David Traum. 2011. Improving Spoken Dialogue Understanding Using Phonetic Mixture Models. *in Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference (FLAIRS-24)*. Palm Beach, Florida, USA 18-20 May 2011.

William Yang Wang and Julia Hirschberg. 2011. Detecting Levels of Interest from Spoken Dialog with Multistream Prediction Feedback and Similarity Based Hierarchical Fusion Learning. *in Proceedings of the 12th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2011)*. Portland, Oregon, USA 17-18 June 2011.

Biographical Sketch

I am currently a first-year PhD student and the R.K. Mellon Presidential Fellow at the Carnegie Mellon School of Computer Science, working with Justine Cassell of the Human-Computer Interaction Institute and Alan W Black of the Language Technologies Institute. I have broad interests in Computational Linguistics, Machine Learning, and Spoken Language Processing.

Previously, I received an M.S. in Computer Science from Columbia, advised by Kathy McKeown and Julia Hirschberg. I publish at conferences and journals such as ACL, COLING, SIGDIAL, INTERSPEECH, ASRU, IJCNLP, and CSL. I also serve as reviewer for journals such as IEEE Transactions on Affective Computing, IEEE Signal Processing Letters and IEEE Communications Letters. I was on the Organizing Committee of the 7th Young Researchers' Roundtable on Spoken Dialog Systems (YRRSDS 2011). I have held visiting researcher positions at the USC Institute for Creative Technology, and Chinese University of Hong Kong (CUHK)/Chinese Academy of Sciences (CAS). I am currently affiliated with Microsoft Research Redmond.


YOUNG RESEARCHERS' ROUNDTABLE
YRR?
SDS!
ON SPOKEN DIALOGUE SYSTEMS