ISSN 2186-7437

NII Shonan Meeting Report

No. 2015-7

The Future of Human-Robot Spoken Dialogue: from Information Services to Virtual Assistants NII Shonan Meeting Report

Rafael E. Banchs
Sakriani Sakti
Etsuo Mizukami

March 26–28, 2015

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo, Japan

The Future of Human-Robot Spoken Dialogue: from Information Services to Virtual Assistants NII Shonan Meeting Report

Organizers:
Rafael E. Banchs (Institute for Infocomm Research, Singapore)
Sakriani Sakti (Nara Institute of Science and Technology, Japan)
Etsuo Mizukami (National Institute of Information and Communications Technology, Japan)

March 26–28, 2015

Meeting Motivation and Overview

Dialogue Systems embrace the ultimate goal of human-robot interaction, in which computational systems communicate with their human interlocutors in the same way humans communicate among them. Although significant progress has been achieved during the last few years and some pioneering commercial systems are already finding their way to the market, current state-of-the-art dialogue systems are very limited in their approach to the human communication phenomenon. Some of these limitations include:

- the lack of ability to properly model world knowledge for reasoning purposes,
- the still low reliability of speech recognition and the limited capacity of dialogue systems to cope with recognition errors in a logical manner,
- the complex and multimodal nature of the pragmatic phenomena and the socio-cultural aspects that affect such interactions, and
- the difficulty of evaluating dialogue quality and the subsequent development of reliable strategies for automatic learning and adaptation.

The main objective of this NII-Shonan meeting was to discuss about the most relevant and promising future directions of research in dialogue systems. The discussion was centered on how these research directions address the different problems and limitations of current dialogue systems, as well as how they provide the basis for the next generation of intelligent artificial agents.

The participants were requested to present work in progress as well as to propose new collaborative efforts to tackle the main limitations of current dialogue systems mentioned above. As a result of the meeting, a research agenda for main directions and international collaboration for the next three to five years was defined.

Joint organization of future workshops and shared tasks by the participants in the meeting were defined with the objective of pushing the state-of-the-art in dialogue systems to a more comprehensive effort for developing a new generation of intelligent virtual assistants.

Seminar Schedule

March 25th (Wednesday): Check-in

15:00- Check-in

19:00- Welcome Banquet

March 26th (Thursday): Day 1 Full-day session (Room 208)

07:30-08:30 Breakfast

08:30-08:35 Seminar Opening

08:35-10:00 Session 1: Social Human-Robot Interaction

10:00-10:30 Break

10:30-12:00 Session 2: Artificial Conversational Agents

12:00-13:30 Lunch

13:30-15:00 Session 3: Knowledge Representation and Dialog Modelling

15:00-15:30 Break

15:30-17:00 Session 4: Spoken Dialog and Paralinguistics

17:00-18:00 Free discussion period

18:00-19:30 Dinner

March 27th (Friday): Day 2 Full-day session (Room 208 & 209)

07:30-08:30 Breakfast

08:30-09:00 Team Formation and Round Table Instructions (Room 208)

09:00-10:00 Session 5: Round Table Discussions (Room 209)

10:00-10:30 Break

10:30-12:00 Session 6: Round Table Discussions (continuation)

12:00-13:30 Lunch

13:30-15:00 Session 7: Plenary Sessions (Part I) (Room 208)

15:00-15:30 Break

15:30-17:00 Session 8: Plenary Sessions (Part II)

17:00-18:00 Free discussion period

18:00-20:00 Main Banquet

March 28th (Saturday): Day 3 Half-day session (Room 208)

07:00-10:00 Check out

07:30-08:30 Breakfast

08:30-09:00 Working Committee Formation

09:00-10:00 Session 9: Working Committee discussions

10:00-10:30 Break

10:30-12:00 Session 10: Road map presentations for selected initiatives

12:00-13:30 Lunch

13:30 Dismiss

Overview of Talks in Session 1: Social Human-Robot Interaction Session Chair: Satoshi Nakamura

Symbiotic human robot interaction

Tatsuya Kawahara, Kyoto University

A new project on symbiotic human robot interaction is introduced. The goal

of the project is autonomous humanoid robot who looks like human, behaves like human, and interacts like human. Key technologies include robust speech recognition, dialogue modelling and multi-modal interaction modelling.

Engagement and natural interaction in human-robot interactions

Kristiina Jokinen, University of Helsinki

In human-human communication, a wide repertoire of multimodal signals is used to provide effective feedback about the parther's interest and understanding. In a similar manner, multimodality is important in the context of "social robotics", where the robot is meant to support human users in social, interactive tasks, and to serve as an intuitive interface to access and share information.

The interlocutors' experience of the interaction is largely based on their holistic interpretation of the partner's multimodal behaviour, besides the content of the verbal communication. We hypothesise that by observing the user's behaviour when interacting with a humanoid robot, it is possible to estimate how a shared understanding is constructed through interaction among the interlocutors, and consequently, to predict the users experience of their interaction with the robot. The users assessment of the interaction (responsiveness, expressiveness, interface, usability, overall impression) indeed seems to correlate with their multimodal behaviour, i.e. the interlocutors engagement and active participation relate to their assessment of the success of communication. This allows us to establish measurements for the users engagement and dialogue management in robot interactions, and points towards automating evaluation of human-robot interactive situations.

Appropriateness of behaviours as a participant of conversations: can a robot participate in the multi-party dialogue?

Etsuo Mizukami, National Institute of Information and Communications Technology

In the near or long-term future, a robot may exist as a member of our society. What a robot should do to be such an existence even in the conversation made by multi-party? In order to answer this question, we have to know how we human behave in the multi-party dialogue. Although a robot need not to behave as same as a human, it is required at least to understand some norms or rules of human interaction. For example, it has to know when it should start to speak relevant words as a proper role, and what it should do as an appropriate hearer.

Human-robot interaction evaluation and ethics

Joseph Mariani, LIMSI-CNRS & IMMI

I coordinated within the CNRS Ethics Committee (Comets) a working group who produced a report on the Ethics of research in ICT. This report contains several recommendations. Generally speaking, the report expressed the fact that in order to be more respectful of ethics, research in ICT should include

a reflection on the consequences of its results at the time when the research is being conducted, and eventually adapt or complete the scope of research. It went with several examples, including one on robotics: one of Asimov laws mentions that a robot shouldnt harm a human. This is only possible if the robot is able to identify humans, and make the distinction between humans and objects, animals or other robots. It therefore implies that the robot possesses vision abilities, which is the scientific field of Computer Vision. The present state-of-the-art in Computer Vision appears however insufficient to guarantee an acceptable categorization by the robot between humans and non-humans. It is therefore necessary to invest on Computer Vision in order to measure the quality of the systems, the progress in quality and the adequacy with the robot needs in order to behave in an ethical way.

The report proposed to launch a specific national Ethics Committee for ICT research, which is now active. The first report of this committee deals with the ethics of research in robotics, and extends very largely the conclusions of the first report. One of the issues that it contains concerns the traceability of the robots actions, and the fact that it should be able to explain clearly and quickly, and make understandable, the reasons of its decisions and actions, based on its perception.

Together with Laurence Devillers who participated in this report on research in robotics, I would like to propose to carry on a reflection within this meeting on the evaluation of human-robot interaction, as a necessary tool in order to ensure the ethical behaviour of robots. This would include the study of the various components and functions that should be evaluated, of the general organization of evaluation, of the various metrics that could be used, either based on human or automatic evaluation, of the production and distribution of development and test data, including the questions of privacy, etc.

Affective and social robotics: engagement, evaluation and ethics

Laurence Y. Devillers, LIMSI-CNRS/Univ. Paris-Sorbonne

I would like to propose to carry on a reflection within the Shonan meeting on advanced dialogues employing complex social behaviours in order to provide a companion-robot with the skills to create and maintain a long term social relationship through verbal and non verbal language interaction. Talk during social interactions naturally involves the exchange of propositional content but also and perhaps more importantly the expression of interpersonal relationships, as well as displays of emotion, affect, interest, etc. Such social interaction requires that the robot has the ability to represent and understand some complex human social behaviour. Cognitive decisions will be used for reasoning on the strategy of the dialog and deciding more complex social behaviours (humour, compassion, white lies, etc.) taking into account the user profile and contextual information. Social interactions require social intelligence, cultural understanding and robust multimodal perception and employ theory of mind for inferring the cognitive states of another person. It is not straightforward to design a robot with such abilities and it is necessary to combine pluri-disciplinary skills and theories.

This reflection will also include the evaluation of such system and the various metrics that could be used like the measure of social engagement with the user. With Joseph Mariani we would like to carry out a discussion on evaluation and also on how researchers could incorporate ethical constraints at the stages of defining and implementing their research projects.

Models of culture-specific dialogue

David Traum, USC ICT

Virtual humans are artificial agents that include both a visual human-like body and intelligent cognition driving action of the body, including engaging in face to face conversation. Culture covers a wide range of common knowledge of behaviour and communication that can be used in a number of ways including interpreting the meaning of action, establishing identity, expressing meaning, and inference about the performer. Virtual human and robot behaviour will always be interpreted by people from a culture-specific vantage point and viewers will make inferences about cultural aspects of the virtual humans, so whether or not an explicit model of culture is used in the design and behaviour of the virtual humans, one will be attributed to them. In this talk, I will present a taxonomy of types of culture models for virtual humans and look at several examples of existing cultural models that have been used, focusing primarily on those we have developed at the Institute for Creative Technologies at University of Southern California, and point out remaining steps for a more full model of culture.

Overview of Talks in Session 2: Artificial Conversational Agents Session Chair: Wolfgang Minker

Multiparty and open domain reasoning dialog system

Gary Geunbae Lee, POSTECH

Will present recent postech approaches for multi-party dialog system and intelligent assistant dialog system which shows almost open domain language understanding and knowledge based reasoning capabilities.

Turn-taking and attention in human-robot dialogue

Gabriel Skantze, KTH

At KTH we have been doing research on multi-modal dialogue systems for a long time, often based on observations of human-human dialogue. In recent years, we have focused on issues related to human-robot dialogue, such as situated interaction, multi-party dialogue and visual attention. We have developed several tools for conducting this research, including the dialogue system framework IrisTK (released as open source) and the back-projected robot head Furhat (which is now a commercial product). I will describe how we have used these

tools for studying how turn-taking and attention is coordinated in situated, multi-party dialogue. This includes controlled experiments in the lab, as well as large scale field trials in public settings.

Towards open-domain conversational systems

Ryuichiro Higashinaka, NTT

Although task-oriented dialogue systems have been actively investigated in the past decades, it is only recently that open-domain conversational systems, or chat systems, have been attracting attention. Such systems need to handle open-domain utterances, which poses many challenges in dialogue systems research. In my presentation, I describe our ongoing work on using predicate argument structures (with zero-anaphora resolution) for open-domain utterance understanding and using social media as resources for natural language generation about a wide variety of topics. I also talk about how to personalize such open-domain conversational systems for long-term use. I briefly describe the project in Japan which aims at detecting dialogue breakdown in open-domain conversation.

A chat agent based on the vector space model

Rafael E Banchs, Institute for Infocomm Research

This presentation will focus on IRIS (Informal Response Interactive System) a chat engine that implements a vector space search approach over a data collection of movie dialogs. I will describe in detail the basic principles of operation of the system, along with the added capabilities of vocabulary learning and style and manner adaptation. A comparative evaluation between the proposed system and a baseline system is also presented. Finally, future plans of research in this area will be shared.

Developing conversational agents and the studies of affective and cognitive processes

Sakriani Sakti, Nara Institute of Science and Technology

Interaction between human and computer continues to change to the better replicate interaction between humans. The aim is to build a conversational agent that can interact with human in as natural a fashion as possible. In this talk, I will present some ongoing research works in developing conversational agents. This includes chat-based dialog management, emotion recognition, paralinguistic expression, and cognitive communication.

CLARA: a virtual agent for conference information and touristic assistance

Haizhou Li, Institute for Infocomm Research

I will present some results from our recent experience on deploying a virtual agent called CLARA at INTERSPEECH 2014. The virtual agent provided con-

ference delegates with information related to the conference's technical program and related events, as well as touristic information about Singapore.

Overview of Talks in Session 3: Knowledge Representation and Dialog Modelling Session Chair: Gary Geunbae Lee

Incorporating knowledge into word representation learning

Kevin Duh, Nara Institute of Science and Technology

Recent work has shown that word vector representations, also known as word embeddings, can successfully capture semantic and syntactic regularities in language (Mikolov2013) and improve the performance of various Natural Language Processing systems (Collobert2011). While many methods have been proposed, most are based on the same premise of "distributional semantics", where words from similar contexts are mapped to nearby vectors.

However, distributional semantics is by no means the only way to model word meaning. In this talk, I will discuss our attempts to incorporate relational semantics and world knowledge into word representation learning algorithms. I will describe a general algorithm based on the Alternating Direction Method of Multipliers (ADMM) that enables us to integrate flexible knowledge sources such as WordNet and FreeBase.

Towards improving virtual assistants intelligence

Jiang Ridong, Institute for Infocomm Research

Speech interface is an intuitive, flexible and natural means of communication between users and machines. With the advancement of natural language processing technology, more and more voice enabled applications and products are emerging in our daily life. One of the great challenges is how to improve the intelligence of these virtual assistants. Various systems with different intelligences have been developed by researchers from all around the world. However, there is still a long way for a machine to pass the Turing test to show its intelligence. This requires long-standing efforts of the research community on this topic.

Knowledge based AI for service robotics

Suraj Nair, TUM Create

In order to systematically introduce robots into human spaces, the usability of such robots needs very close attention. Human with no knowledge in robotics should be able to command and interact with such systems in a intuitive and ergonomic manner. The main challenge here is to build system which is capable of receiving under-specified instruction from a human and still be able to execute the tasks successfully. In order to achieve this we propose a Knowledge Based system where domain knowledge is expressed in a semantic form. The challenge here is to model common sense knowledge and logical reasoning tools which can

infer missing pieces of information important for task execution. We will discuss our efforts in this direction and current results.

Dialog state tracking on human-human dialogs

Seokhwan Kim, Institute for Infocomm Research

Dialog state tracking is one of the key sub-tasks of dialog management, which defines the representation of dialog states and updates them at each moment on a given on-going conversation. Most previous work on dialog state tracking has focused on developing trackers based on given data which consist of goal-oriented human-machine conversations for searching bus schedules, restaurant, or tourist information. Alternatively, we suggest a new dialog state tracking task on human-human dialogs as a shared task. Although it would be difficult to apply the outputs from this task immediately to any practical systems, we expect this as a first step to develop much more human-like systems in a long-term point of view.

WFSD based multilingual spoken dialogue management for human-robot communication

Takuma Okamoto, National Institute of Information and Communications Technology

For realizing human-robot spoken dialogue systems, not only one user situation but also multiuser situation should be considered. In addition, in globalized society, multilingual spoken dialogue systems are required so that a robot can speak multilingually for multinational uses simultaneously. The WFST based spoken dialogue manager can realize these multilingual spoken dialogue systems by unifying each spoken language understanding to all together as a multilingual spoken language understanding WFST.

Dialog control considering systems goal and incongruity detection by brain signals

Satoshi Nakamura, Nara Institute of Science and Technology

My presentation includes 1) the current dialog paradigm needs be more generalized so that the system can persuade users or negotiate with users, 2) the dialog control needs to consider paralinguistic information like emotions, 3) realtime objective evaluation methods needs to be developed, and 4) the persona for the dialog system needs to be considered.

Overview of Talks in Session 4: Spoken Dialog and Paralinguistics Session Chair: Kristiina Jokinen

Conversational interfaces; issues and technologies

Michael McTear, University of Ulster

My current research is in the area of conversational interfaces. We are looking at how people will be able to communicate with devices such as smartphones, robots, smart watches, and other wearables. A colleague and I are currently writing a book on this topic and one of our main tasks is to define the issues involved and technologies required to address them. One of the main challenges concerns the interpretation of the users input and the determination of the devices response. Identifying the users intent involves a combination of dialogue act recognition and topic identification for example, determining whether the user is requesting information, some action, or just making conversation. Given that devices are connected and that they can provide useful information about the conversational context, the user location, as well as information provided by biosignals, the context for interpretation is much wider than in traditional spoken language understanding. Next the system has to decide the appropriate action to take for example, an Internet search, a physical action, a conversational turn and which device or service is involved. Finally the response has to be formulated, for example, as a combination of speech and action, a summary from a web document, or a multimodal output. Although these issues have been investigated extensively in the different technologies that contribute to spoken dialogue systems - spoken language understanding, dialogue management, and response generation -, their application in the area of conversational interfaces to multiple connected devices provides many new challenges and opportunities.

Interaction with cognitive technical systems

Wolfgang Minker, Ulm University

Spoken Language Dialogue Systems (SLDSs) providing natural interfaces to computer-based applications have been an active research area for many years. However, most state-of-the-art systems are still static in use, and their field of application is rather limited. Our work aims at overcoming this limitation. In this presentation, we will provide a brief overview on our research activities in the domain of adaptive and assistive SLDSs for next-generation Cognitive Technical Systems. These activities include Emotion Detection from Multimodal Signals, Verbal Intelligence Estimation based on the analysis of Spoken Utterances and Statistical Modelling for User-centred Adaptive Spoken Dialogue Systems.

Spoken dialog technologies in NICT

Atsuo Hiroe, National Institute of Information and Communications Technology

This presentation introduces spoken dialog technologies developed in Na-

tional Institute of Information and Communications Technology, NICT. One of issues in a spoken dialog system is the trade-off between accuracy of its response and portability of language or domain. In order to address this issue, we represent both a model for language understanding and a dialog scenario as a finite state transducer (FST) and drive it with an FST decoder. This framework can eases the trade-off since FST has potential to express both a rule and a statistical model while FST decoder is language and domain independent. We also introduce actual porting works.

Flexible speech generation and input technologies for making man-machine communication more effective

Tomoki Toda, Nara Institute of Science and Technology

Speech can convey not only linguistic information but also para-linguistic information and non-linguistic information simultaneously. To make man-machine communication more effective, it is useful to develop technologies for handling para-/non-linguistic information as well, e.g., a speech generation technique to make it possible to flexibly control para-/non-linguistic information. Moreover, it will also helpful for making man-machine communication available anytime and anywhere to develop technologies for achieving more flexible speech input, such as a silent speech technique for privacy purpose. In this talk, I will introduce our recent work on the development of flexible speech generation and input technologies.

Help your robot understand you: cooperative learning in computational paralinguistics

Björn Schuller, University of Passau & Imperial College London

Human-Robot Spoken Dialogues to the present day mostly lack emotional and social intelligence. This can be overcome to some degree by lending our new electronic companions the ability to analyse our state and trait in a reliable and robust manner. In this light, the state-of-the-art in computational paralinguistic speech analysis for robots is highlighted based on the Interspeech ComParE 2009-2015 series initiated and co-organised by the presenter. To improve on these results, avenues towards largely autonomous learning abilities of a robot without the need to permanently require "more input" from the user will be discussed. To this end, the novel cooperative learning approach is presented as an utmost efficient blend of semi-supervised and active learning. The required confidence measurement is quickly introduced as well. The talk is rounded up by a more general view on current activities of the speakers two groups at the University of Passau/Germany and the Imperial College London/UK in this field including big-data deep recurrent memory-enhanced multi-task learning, transfer learning, and distributed processing.

Paralinguistic processing for dialogue systems

Nick Campbell, Trinity College Dublin, The University of Dublin

Speech recognition and synthesis are now mature technologies but are still

predominantly text-based; they are not sufficiently tuned to the characteristics of spoken language. In particular, they typically lack the ability to process the social signals that accompany propositional content in spoken interactions. At TCD we have been building dialogue systems that present solutions to this problem; the incorporate non-verbal information (both speech and image) alongside more traditional dialogue information and have been testing these devices in robotics, call-centre, teaching, and elderly-care applications.

Overview of Round Tables and Working Committee Discussions

During the second and third day of the meeting, Round Table and Working Committee discussions were conducted respectively. The five selected topics and their corresponding Round Table and Working Committee members were as follows:

- Group 1: Multimodality and Social Interaction
  Participants: Tatsuya Kawahara, Kristiina Jokinen, Laurence Devillers, Etsuo Mizukami, Gabriel Skantze, Gary Geunbae Lee
- Group 2: Automatic Speech Recognition, Natural Language Understanding and Natural Language Generation
  Participants: Michael McTear, Seokhwan Kim, Atsuo Hiroe, Björn Schuller, Tomoki Toda
- Group 3: Knowledge Representation and Reasoning
  Participants: Kevin Duh, Jiang Ridong, Suraj Nair, Sakriani Sakti, Haizhou Li
- Group 4: Dialogue Management Paradigms and Policy Learning
  Participants: Satoshi Nakamura, Nick Campbell, Takuma Okamoto, David Traum
- Group 5: Automatic Evaluation and Resources
  Participants: Ryuichiro Higashinaka, Joseph Mariani, Wolfgang Minker, Rafael Banchs

The main objective of Round Table discussions was to focus on specific topics to identify the main challenges currently faced by research work in each area and define specific actions (i.e. workshops, exploratory project, shared task or competitions, open source development, etc.) to be proposed to the community. The main objective of the Working Committee discussions was to select specific actions proposed during the previous day and generate: an actiona plan for the selected action, a tentative timeline of activities, and a list of required resources and potentially interested teams.

List of Participants and Affiliations

The following is a complete list of participants (listed in no particular order) and their affiliations:

- Prof. Satoshi Nakamura, Nara Institute of Science and Technology, Japan
- Prof. Tatsuya Kawahara, Kyoto University, Japan
- Dr. Ryuichiro Higashinaka, NTT, Japan
- Prof. Gary Geunbae Lee, Postech, South Korea
- Dr. Haizhou Li, Institute for Infocomm Research, Singapore
- Dr. Suraj Nair, TUM Create Singapore / TUM Germany, Germany
- Prof. Wolfgang Minker, Ulm University Communications Engineering, Germany
- Prof. Michael McTear, University of Ulster, UK
- Prof. Nick Campbell, Trinity College, Dublin, Irland
- Prof. Kristiina Jokinen University of Helsinki / University of Tartu, Finland
- Prof. Joseph Mariani, LIMSI-CNRS & IMMI, France
- Dr. David Traum, USC ICT, USA
- Prof. Laurence Devillers, LIMSI-CNRS/University Paris-Sorbonne, France
- Assoc. Prof. Tomoki Toda, Nara Institute of Science and Technology, Japan
- Assis. Prof. Kevin Duh, Nara Institute of Science and Technology, Japan
- Dr. Seokhwan Kim, Institute for Infocomm Research, Singapore
- Dr. Ridong Jiang, Institute for Infocomm Research, Singapore
- Prof. Björn Schuller, Imperial College London / University of Passau, UK
- Assis. Prof. Gabriel Skantze, KTH Royal Institute of Technology, Sweden
- Dr. Takuma Okamoto, National Institute of Information and Communications Technology, Japan
- Mr. Atsuo Hiroe, National Institute of Information and Communications Technology, Japan
- Dr. Rafael E. Banchs, Institute for Infocomm Research (I2R), Singapore
- Assis. Prof. Sakriani Sakti, Nara Institute of Science and Technology (NAIST), Japan
- Dr. Etsuo Mizukami, National Institute of Information and Communications Technology (NICT), Japan