AIST

# Dialogues, Data and Daily Activities
# -
# Research on Socially Intelligent Robots

Kristiina Jokinen
AI Research Center
AIST Tokyo Waterfront

January 15, 2019



AIST

## Human-Robot Dialogue Systems

- AIRC RobotTalk (Jokinen et al. 2018)
  - Basic care-taking tasks
- WikiTalk, MoroTalk, SamiTalk (Wilcock and Jokinen 2013)
  - Open domain multilingual dialogues from Wikipedia
- Android ERICA (Ishiguro et al. 2012; Kawahara et al. 2017)
  - Multimodal dialogues with a human-like robot
- Intelligent speakers – not robots !
  - Chat-bots with no capability to move or have multimodal dialogues

AIST

# Ability to Communicate

- Collaboration requires communication
  - Participants have mutual knowledge and share an interpretation framework
- Knowledge of the context and reasoning
  - If the human asks a robot “give me the bowl” which physical entity is referred to?
- Grounding is essential:
  - Anchor language symbols to perception (vision)
  - Confirm with the partner of the referents for the used words (mutual knowledge)
- Different levels of interaction
  - With the environment (lights come on), objects (mobile phone),
  - With humans (language-based communication), intelligent agents (robots)

AIRC

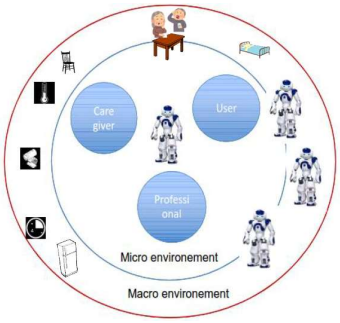


3

AIST

# Robot’s two roles in HRI

- Robot as a computer and a tool
  - Human control, transparent actions
  - Knowledge sources
    - Database, internet, environment, partner
    - Planning and plan execution
  - Knowledge representation
    - Symbols, vectors, actions
    - Representation learning and complex reasoning




- Robot as an agent
  - Language and conversational structure create expectations of social interaction as opposed to interaction with a tool (Jokinen 2009)
  - Can create attachment, therapeutic bond (Bickmore et al. 2005)
  - Media equation (Nass and Reeves 1996)

AIRC

Jokinen, K. (2018). Dialogue Models for Socially Intelligent Robots. The 10th International Conference on Social Robotics (ICSR), Qingdao, China.

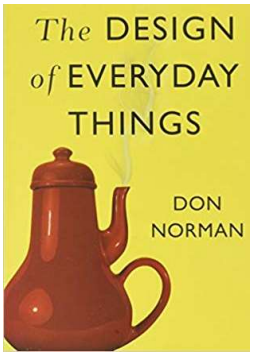


4

AIST

# Interaction Affordances

- Vision: object and their functions (Gibson 1979)
- Usability: The Design of Everyday Things (Norman 1986)
- Robotics: action possibilities (Marin-Urias et al. 2009)
- Interactive systems: natural communication possibilities (Jokinen 2009)
  - adopt models and vocabulary that are well suited for machine processing
  - remain as close as possible to the human's own communication structures and vocabulary
  - use non-verbal forms of communication, like gaze and gesturing




Jokinen, K. (2010). Rational Communication and Affordable Natural Language Interaction for Ambient Environments. In: G.G. Lee et al. (Eds.) IWSDS 2010, LNAI 6392, pp. 163-168. Springer-Verlag, Berlin.

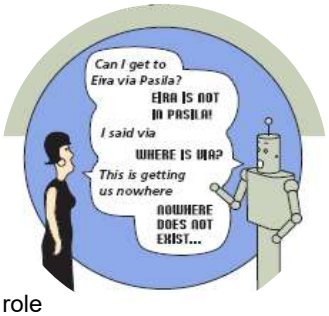


5

AIST

# Social Robots Require Situational Awareness

- Knowledge of what is going on around the agent
- Level of the robot's autonomous behaviour
- Knowledge of the world
- Attention in human conversations
- Feedback: awareness is communicated to partner
  - Intention
  - Engagement
  - Attention
- Eye-gaze in human and agent interactions plays a role
  - Shared attention
  - Turn-taking
  - Feedback
  - Build trust and rapport
- Constructive Dialogue Modelling (Jokinen 2009)



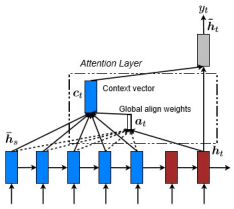



6

AIST

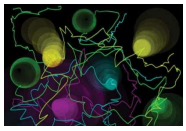
# Attention

- Deep learning
  - When processing a sequence, the network is forced to focus on different parts of the sequence unevenly
  - E.g. different words are important when seen as part of a wider context



Luong et al. 2015

- Visual attention:
  - Human cognitive process through which we get input from the surrounding world
  - E.g. humans need to learn to pay attention to intonation, gaze, gestures, motion, changes in the environment


Findlay and Gilchrist 2012

*In this talk we apply attention to the interaction as a whole (= what are the important elements that the speaker pays attention to when conversing), not just to an input sequence*



AIST

# Eye-tracker studies

- Does eye-gaze help in predicting turn-taking possibilities?
  - mutual gaze to agree to change turns; hesitation pauses gaze aversion

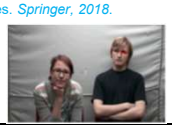
*Jokinen, K., Furukawa, H., Nishida, M., Yamamoto, S. (2013). Gaze and Turn-taking behaviour in Casual Conversational Interactions. ACM Transactions on Interactive Intelligent Systems (TiiS) Journal, Special Section on Eye-gaze and Conversational Engagement, Guest Editors: Elisabeth André and Joyce Chai. Vol 3, Issue 2.*

- Does a silent partner's non-verbal activity influence the other participants' gaze behavior in a three-party conversation situation?
  - Longer fixations not often (to silent partner), more short fixations (to active partner)

Levitski, A., Radun, J., Jokinen, K. (2012). Visual Interaction and Conversational Activity. *The 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Eye Gaze and Multimodality*, at the 14th ACM International Conference on Multimodal Interaction. ICMI.

- Does the speaker's language skills affect communication and eye-gaze?
  - more to the speaker and the speaker's mouth

Ijuin,K., Umata, I., Kato, T., Yamamoto, S. (2018). Eye-gaze and floor apportionment in L1 and L2 dialogues. *Springer, 2018.*

AIST

# Eye-gaze and Intelligent Agents

- Human gaze in different contexts
  - Vertegaal et al (2003): Video-conferencing
  - Gullberg & Holmberg (2006): gestures and eye-gaze
  - Jokinen & al. (2012, 2013): turn-taking
  - Endrass, et al. 2009.: cultural differences
- Gaze-model for believable virtual agents
  - Lee et al. (2007): Gaze model for a Virtual Human
  - Sidner et al. (2005): gaze modelling for conversational engagement
  - Nakano and Nishida (2007): eye-gaze model to ground information in interactions with embodied conversational agents
- Can eye-gaze patterns provide information about understanding



9

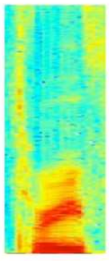
AIST

# Attention: Multimodal Information

Dialogue Summarization

| Start (s) | End (s) | Summarization |
|---|---|---|
| 13 | 18 | greetings, introductions |
| 18 | 27 | occupations, studying language technology |
| 27 | 64 | language technology at the university |
| 64 | 200 | specializing fields in LT |
| 200 | 278 | spoken dialogue systems |
| 278 | 328 | morphological analysis and synthesis, topics in LT |

Acoustic features



Video/Kinect Analysis


Topic Cluster

Laughter annotation

Speaker Movement

Jokinen, K., Trung, NT. (2018). Laughter and Body Movements as Communicative Actions in Interactions. LREC-AREA 2018, Miyazaki, Japan.
Jokinen, K., Trung, NT., Wilcock, G. (2016). Body Movements and Laughter Recognition: Experiments in First Encounter Dialogues. ACM ICMI Workshop MA3HMI'16 , November 16 2016, Tokyo, Japan.



10

AIST

# Encoder-decoder –based video description

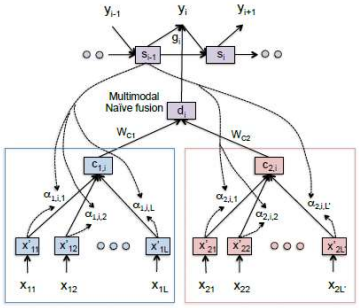



**Image Description to Video Description:**
input is a single static image, output is a sentence

**Add attention** (Xu et al 2015):
- focus on specific parts of the image when generating each word in the description

**Add multimodal attention mechanism** (Hori et al 2017):
- selectively attend to different input modalities (speech and image feature types) and to different times in the input video

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, 2015, pp. 2048–2057.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R. Hershey, Tim K. Marks, and Kazuhiko Sumi, "Attention-based multimodal fusion for video description," in The IEEE International Conference on Computer Vision (ICCV), Oct 2017.

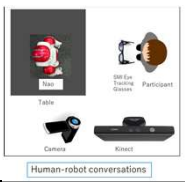


AIST

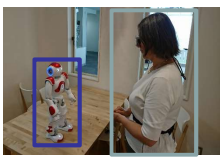
# AIRC Multimodal Data

Each participant had two dyad conversations, one with human and one with robot
Differences of human gaze patterns along:

- Dialogue activity:
  - Instruction giving situations: give structured information related to a task
  - Story-telling: exchange information based on interests, relaxed settings
- Dialogue partner:
  - Human-human and human-robot interactions
  - Compare issues in understanding, misunderstanding, non-understanding
- Language and culture:
  - Japanese – more backchannelling
  - English
- Gender
- Experience with speaking agents, computing







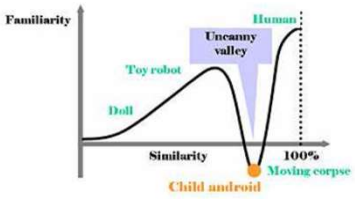
AIST

# Socio-technical systems

*Robots must operate as boundary-crossing agents that facilitate interaction and mutual intelligibility between the perspectives*

=>

*We need to find novel ways to interact with robots as cooperative agents*




AIRC



13

AIST

AIRC

# Thank you!



14