NATURAL LANGUAGE IN HUMAN-ROBOT INTERACTION

Rafael E. Banchs, Seokhwan Kim, Luis Fernando D'Haro, Andreea I. Niculescu

Human Language Technology (I^2R, A*STAR)

HAI 2016
4th International Conference on Human-Agent Interaction
4 - 7 OCTOBER
SINGAPORE

A*STAR Institute for Infocomm Research

About this Tutorial

Tutorial Objective:

- To present a comprehensive description of natural language interaction technologies and its current and potential uses in human-robot interaction applications.

The Speakers:

- Rafael E. Banchs , Seokhwan Kim, Luis Fernando D'Haro, Andreea I. Niculescu
- Dialogue Technology Lab, Human Language Technology, Institute for Infocomm Research (I^2R)

Additional Information:

- From 9:00 to 12:30 with tea break from 10:00 to 10:30
- Feel free to interrupt for clarification questions during the talks. (However, very interesting discussions should be reserved for the tea break or the end of the tutorial.)
- Slides are available at: http://hai-conference.net/hai2016/wp-content/uploads/2016/10/HRI_Tutorial_HAI2016.pdf

Tutorial Content Overview

1. Natural Language in Human-Robot Interaction
   - *Human-Robot Interaction*
   - *The Role of Natural Language*
2. Semantics and Pragmatics
   - *Natural Language Understanding*
   - *Dialogue Management*
3. System Components and Architectures
   - *Front-end System Components (Interfaces)*
   - *Back-end System Components*
4. User Experience (UX) Design and Evaluation
   - *UX Design for Speech Interactions*
   - *User Studies and Evaluation*

PART 1: NATURAL LANGUAGE IN HUMAN-ROBOT INTERACTION

Rafael E. Banchs, Seokhwan Kim, Luis Fernando D'Haro, Andreea I. Niculescu

Human Language Technology (I^2R, A*STAR)

HAI 2016
4th International Conference on Human-Agent Interaction
4 - 7 OCTOBER
SINGAPORE

A*STAR Institute for Infocomm Research

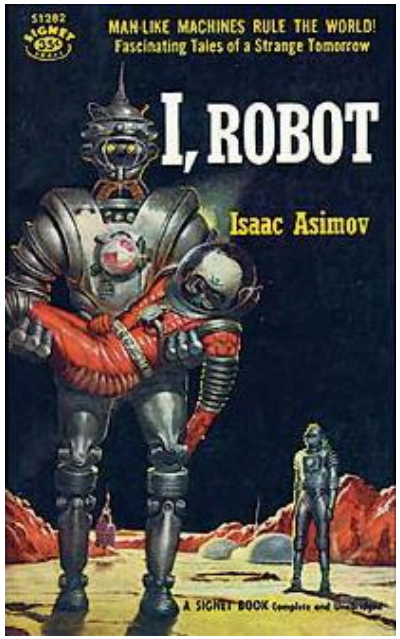
Part 1:
Natural Language in Human-Robot Interaction

HUMAN-ROBOT INTERACTION

Institute for Infocomm Research
A*STAR

What is Human-Robot Interaction About?

- "Human—Robot Interaction (HRI) is a field of study dedicated to understanding, designing, and evaluating robotic systems for use by or with humans" *
- The Three Laws: **
  - *A robot may not injure a human being, or through inaction, allow a human being to come to harm*
  - *A robot must obey the orders given it by human beings except where such orders would conflict with the First Law*
  - *A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws*

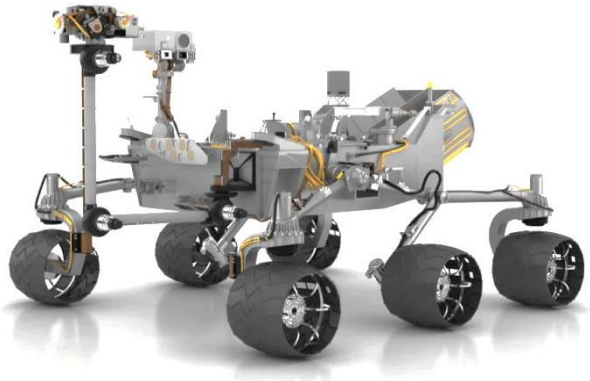


* M. A. Goodrich and A. C. Schultz (2007) Human-Robot Interaction: A Survey. *Foundations and Trends in Human-Computer Interaction* 1(3): 203-275

** Asimov, Isaac (1950). I, Robot.

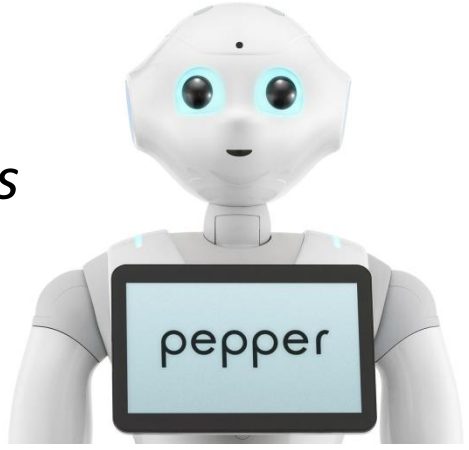
Different Robot Roles and HRI

- **Robots in the wild**
  - *High level of autonomy required*
  - *Multiplicity of functions and resources*
  - *Remote and limited HRI*



- **Robots in the industry**
  - *Low level of autonomy required*
  - *Specificity of functions (controlled and structured environments)*
  - *Programming or command-based HRI*

- **Robots in the society**
  - *Robots as service providers*
  - *Operate in human environments*
  - *Intermediate level of autonomy*
  - *Rich and complex HRI*



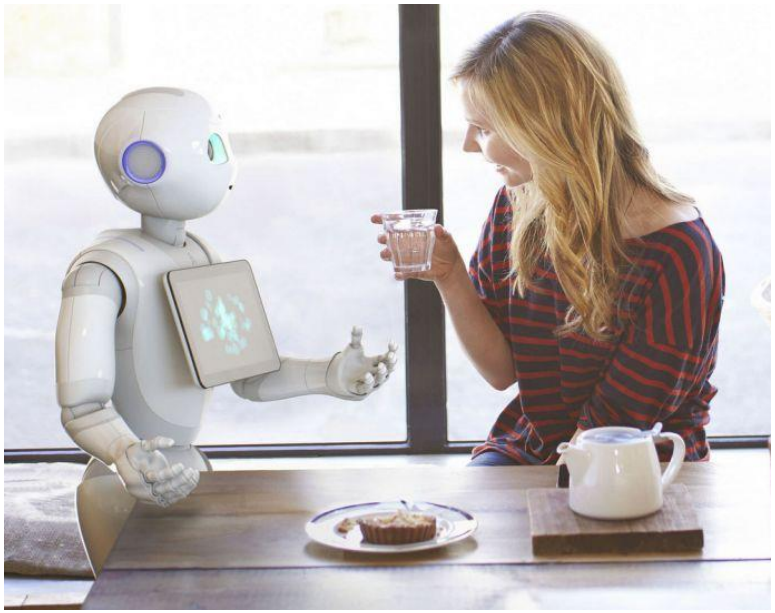
Robots in the Society: Service Robots

- Some roles of Robots in Human Society include*
  - Robots as autonomous machines operating without direct human control (e.g. autonomous vehicles)
  - Robots as sophisticated tools used by a human operator (e.g. medical robotic devices)
  - Robots as participating members of human-centred environments (e.g. social robots, receptionist)
  - Robots as persuasive agents for influencing people's behaviours (e.g. sale robots, therapy robots)
  - Robots as social mediators between people (e.g. language interpreter)
  - Robots as model social actors (e.g. virtual tutor)

* K. Dautenhahn (2003) Roles and Functions of Robots in Human Society - Implications from Research in Autism Therapy. *Robotica* 21(4): 443-452.

The Multimodal Nature of HRI

- Service robots are immerse in the physical world and share spaces with humans
- Multimodality is desirable and needed for HRI with service robots:
  - *Speech and Language*
  - *Vision and Image Analysis*
  - *Tactile Interaction*
  - *Localization, Navigation and Manipulation*
  - *Social Skills**



* David Nield, Robots need manners if they're going to work alongside humans, accessed online at: http://www.techradar.com/news/world-of-tech/future-tech/robots-need-manners-if-they-re-going-to-work-alongside-humans-1304797

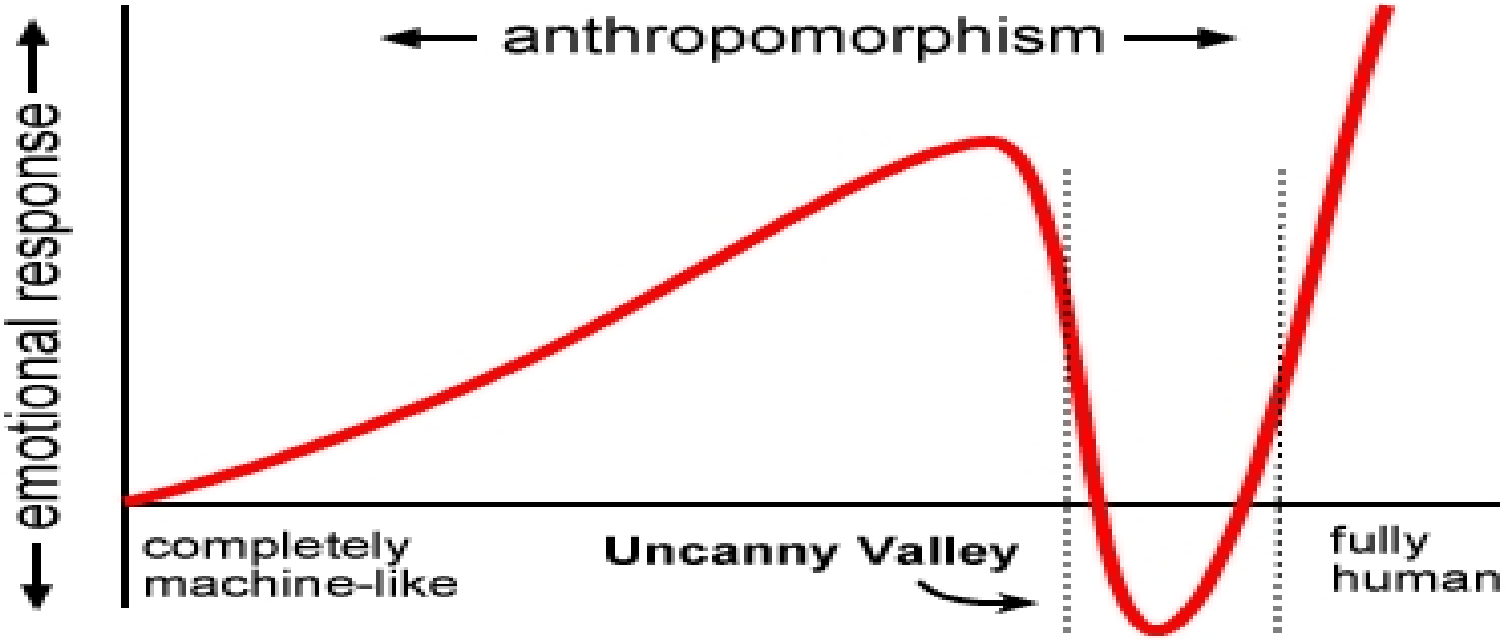
HUMAN-ROBOT VS. HUMAN-HUMAN INTERACTION

- HRI is a synthetic science, not a natural science*
- Some important methodological issues:*
  - The concept of 'robot' is actually a moving target.
  - HRI suffers from not being able to directly compare results from studies using different types of robots.
- Human-Robot Interaction is by definition non-natural:
  - How should non-humanoid robots interact with humans?
  - How much should humanoid robots behave like humans in HRI?

* Kerstin Dautenhahn, Human-Robot Interaction, in The Encyclopedia of Human-Computer Interaction, 2nd Ed., accessed online at https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction

THE UNCANNY VALLEY

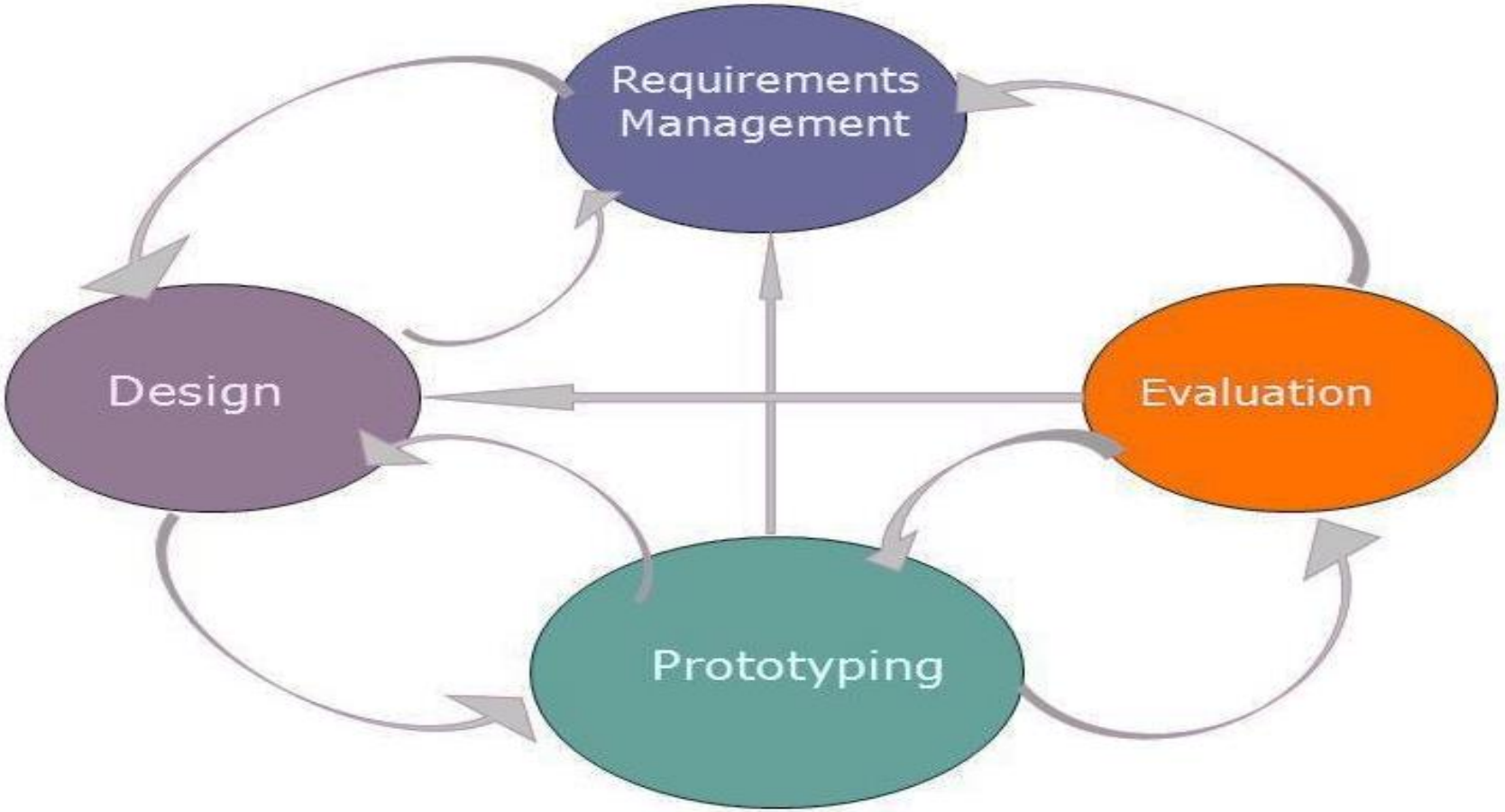
- Area of repulsive response caused by a robot with appearance in between a barely-human and a fully-human entity*



* Mori, M. (2012). The uncanny valley (K. F. MacDorman & Norri Kageki, Trans.). IEEE Robotics and Automation, 19(2), 98–100. (Original work published in 1970).

Interaction Design Lifecycle Model

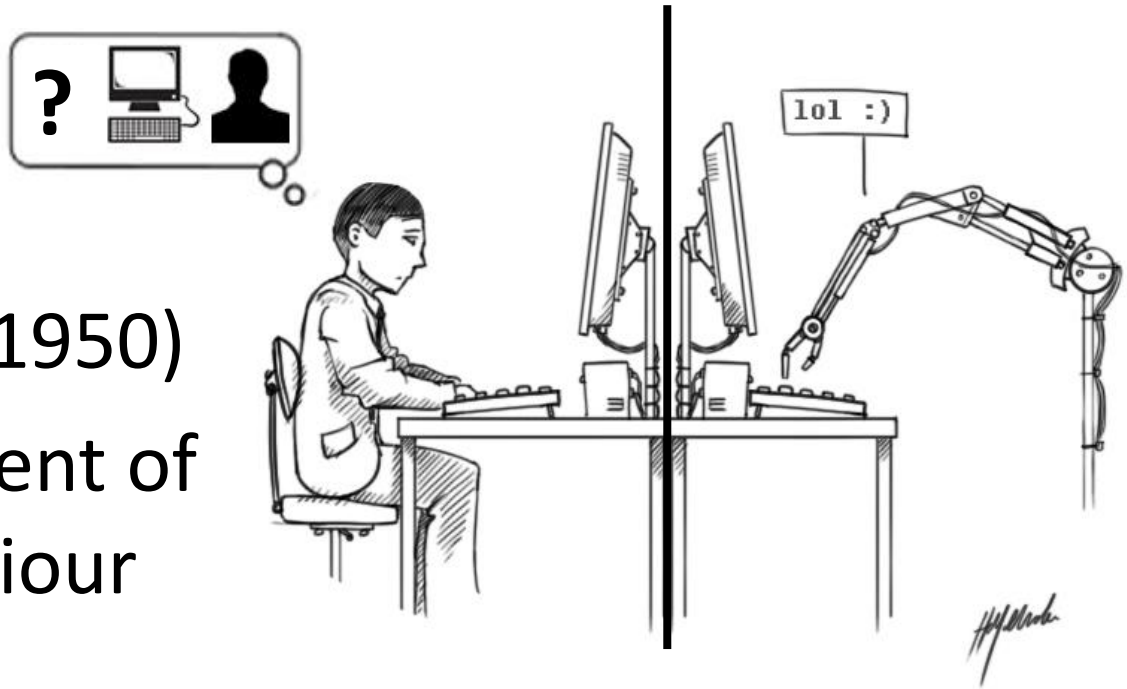
- It is about TECHNOLOGY and **DESIGN**



Taken from: 07 HCI - Lesson 2 Interaction. 07 Outline What is Interaction Design – A multidisciplinary field Terminology – Interaction "Metaphors" and "Paradigms", http://slideplayer.com/slide/4852584/

Artificial Intelligence

- Can robots understand language?
- Can robots actually think?
- Not clear definition of intelligence or how to measure it!



- The Turing Test (1950)
- Indirect assessment of intelligent behaviour

(Image adapted from: http://www.clubic.com/mag/culture/actualite-751397-imitation-game-alan-turing-pere-informatique.html)

Main Challenges for Artificial Intelligence

- Knowledge Representation
  - about learning, storing and retrieving relevant information about the world and one's previous experiences
- Commonsense reasoning*
  - about using world knowledge for interpreting, explaining and predicting daily life events and outcomes

* Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. Commun. ACM 58, 9 (August 2015), 92-103. DOI: http://dx.doi.org/10.1145/2701413

Main References

- M. A. Goodrich and A. C. Schultz (2007) Human-Robot Interaction: A Survey. *Foundations and Trends in Human-Computer Interaction* 1(3): 203-275
- Asimov, Isaac (1950). I, Robot
- K. Dautenhahn (2003) Roles and Functions of Robots in Human Society - Implications from Research in Autism Therapy. *Robotica* 21(4): 443-452.
- David Nield, Robots need manners if they're going to work alongside humans, accessed online at: http://www.techradar.com/news/world-of-tech/future-tech/robots-need-manners-if-they-re-going-to-work-alongside-humans-1304797
- Kerstin Dautenhahn, Human-Robot Interaction, in *The Encyclopedia of Human-Computer Interaction*, 2nd Ed., accessed online at https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/human-robot-interaction
- Mori, M. (2012). The uncanny valley (K. F. MacDorman & Norri Kageki, Trans.). *IEEE Robotics and Automation*, 19(2), 98–100. (Original work published in 1970).
- Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58, 9 (August 2015), 92-103. DOI: http://dx.doi.org/10.1145/2701413

Additional References

- HCI - Lesson 2 Interaction. 07 Outline What is Interaction Design – A multidisciplinary field Terminology – Interaction “Metaphors” and “Paradigms”, http://slideplayer.com/slide/4852584/
- D. Feil-Seifer and M. J. Matarić (2009) "Human-robot interaction ", Invited contribution to *Encyclopedia of Complexity and Systems Science*, pp. 4643-4659, available online at http://robotics.usc.edu/publications/media/uploads/pubs/585.pdf
- Turing, Alan (October 1950) "Computing Machinery and Intelligence", *Mind*, LIX (236): 433–460, doi:10.1093/mind/LIX.236.433, ISSN 0026-4423
- Audrey Oeillet (2015) Imitation Game : ce qu'il faut savoir sur Alan Turing*, le père de l'informatique*, available online at http://www.clubic.com/mag/culture/actualite-751397-imitation-game-alan-turing-pere-informatique.html
- Human-Robot Interaction, A Research Portal for the HRI Community, http://humanrobotinteraction.org/
- HRI Conference, http://humanrobotinteraction.org/category/conference/

Part 1:
Natural Language in Human-Robot Interaction

THE ROLE OF NATURAL LANGUAGE

A*STAR Institute for Infocomm Research

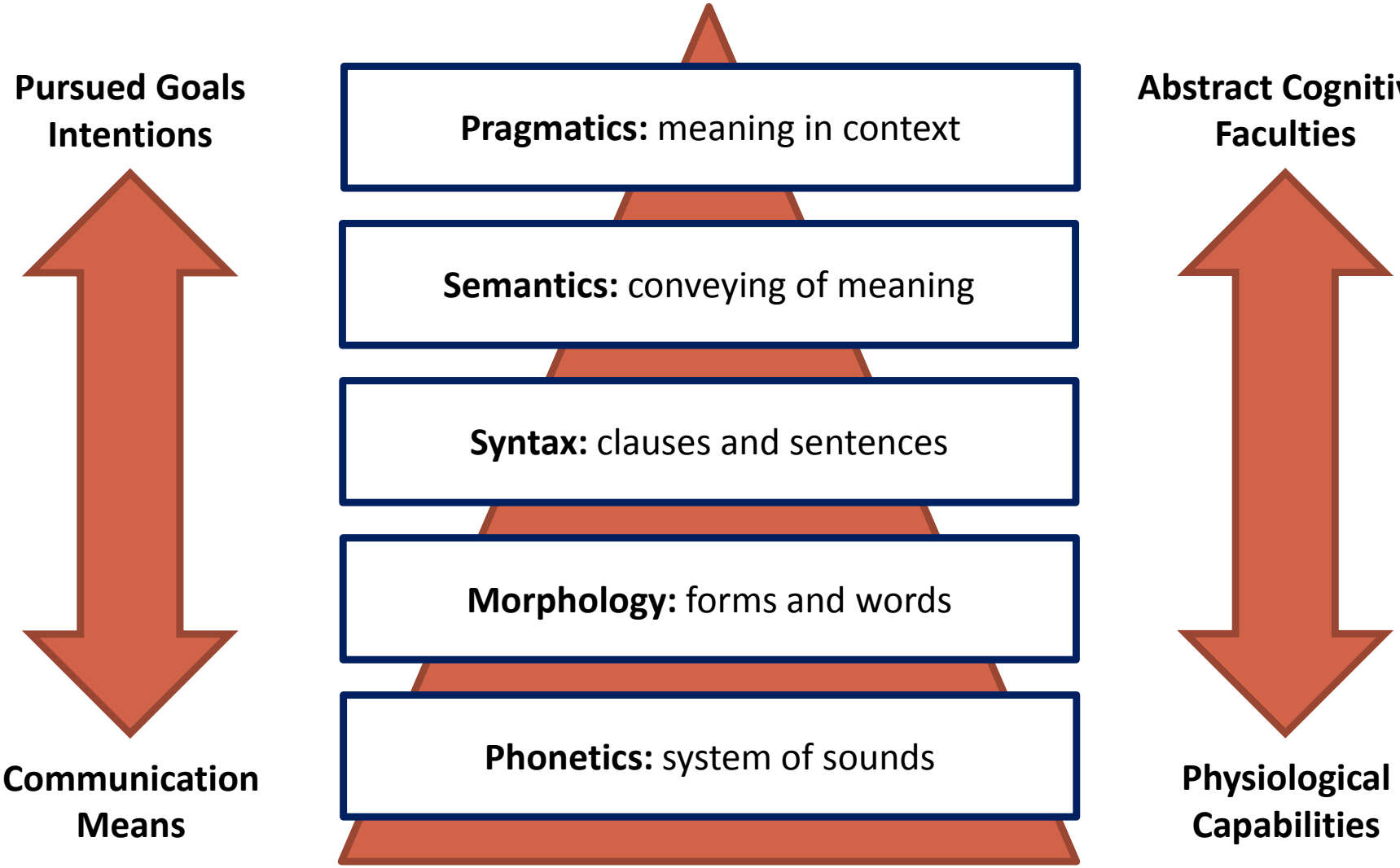
TEACHING ROBOTS TO USE NATURAL LANGUAGE

"The state of the art in natural language interaction allows usable Spoken Dialogue Systems to be developed for robots, that advance beyond simple stand-alone commands."*

- Natural language and speech constitute the most common form of human communication
- Complex tasks require much more than single query or commands
- Dialogue allows for better:
  - *information interchange,*
  - *complex task execution, and*
  - *collaborative work coordination.*
- Robots should learn how to talk to people rather than the other way around!

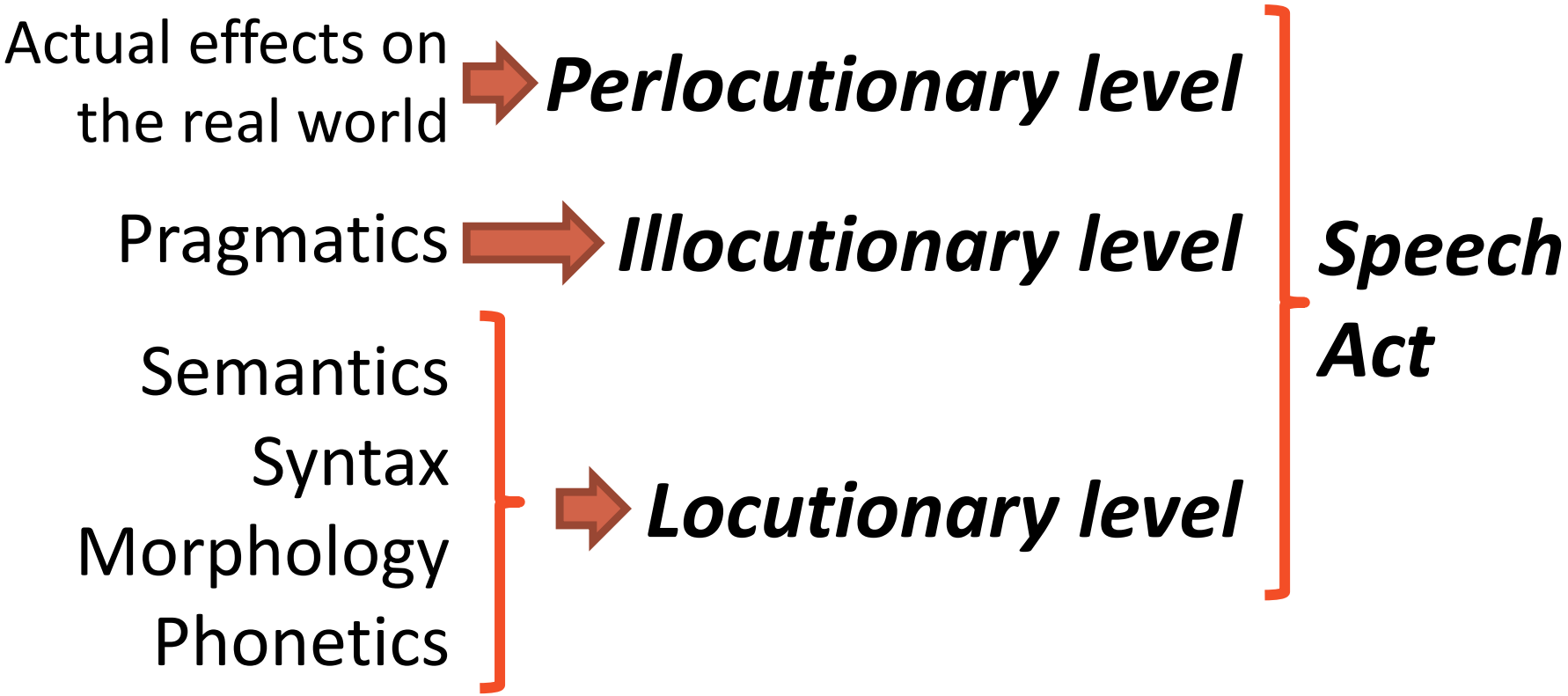
* D. Spiliotopoulos, I. Androutsopoulos, C. Spyropoulos "Human-robot interaction based on spoken natural language dialogue" *Eur. Workshop Service and Humanoid Robots (ServiceRob)* pp. 1057-1060. http://www.aueb.gr/users/ion/docs/servicerob_paper.pdf

Different Levels of the Linguistic Phenomena



The Speech Act Theory

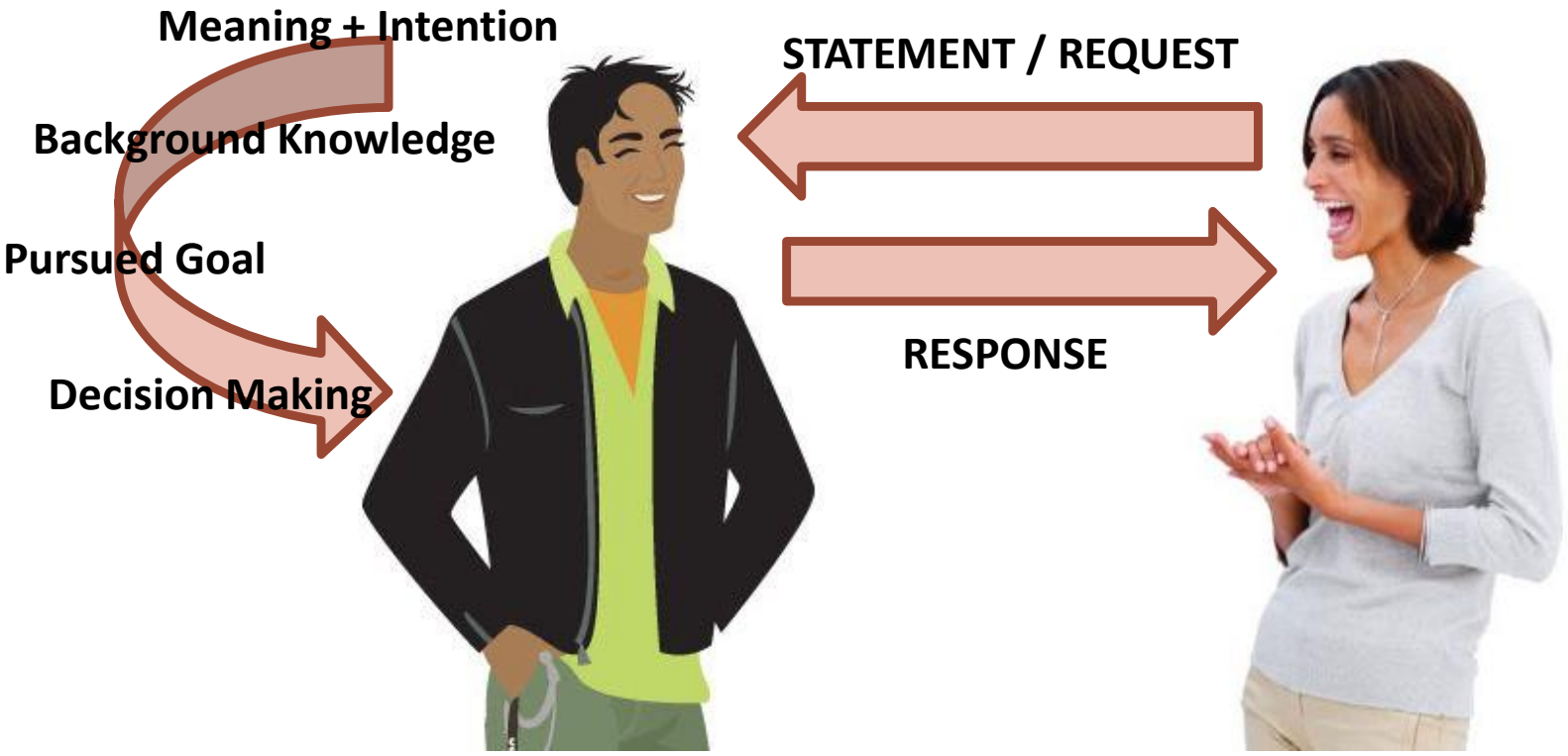
- Dialogue utterances seen as actions taken by the interlocutors



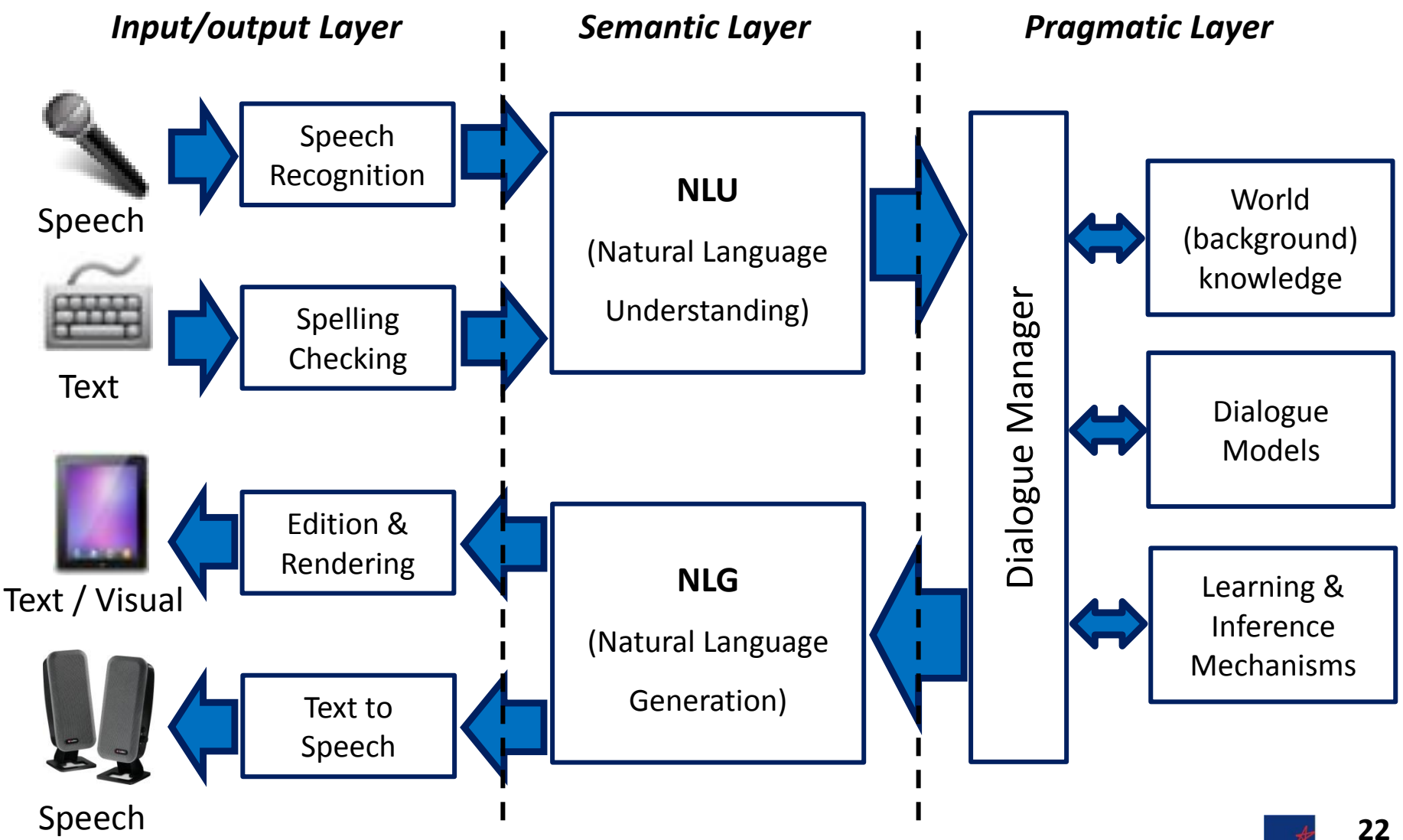
* Austin, J. L. 1962. "How to do things with words". London: Oxford University Press

Collaborative Work among Interlocutors

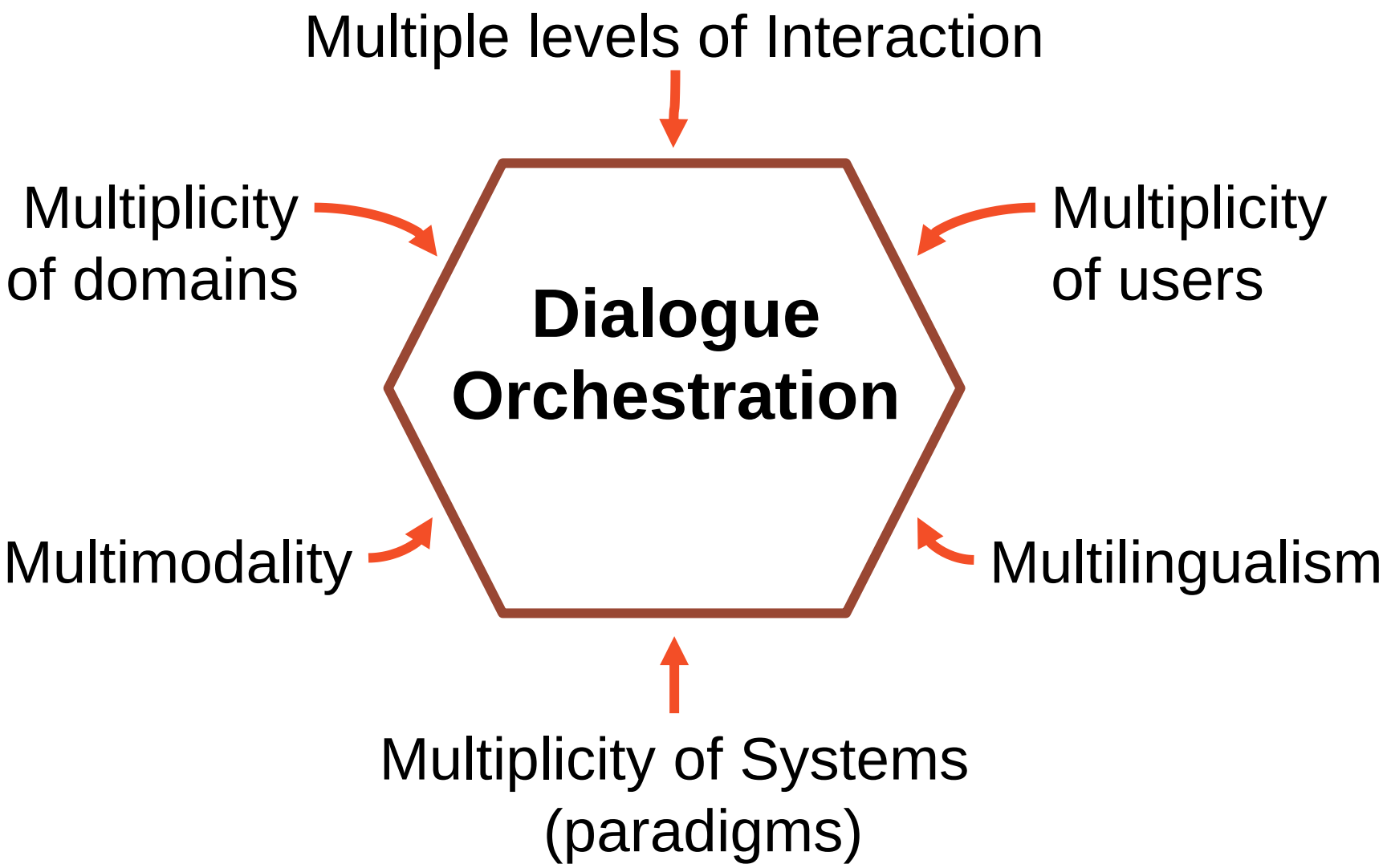
- Dialogue process seen as collaborative work between the interlocutors
- Dialogue management as decision making



General Overview of a Dialogue Engine



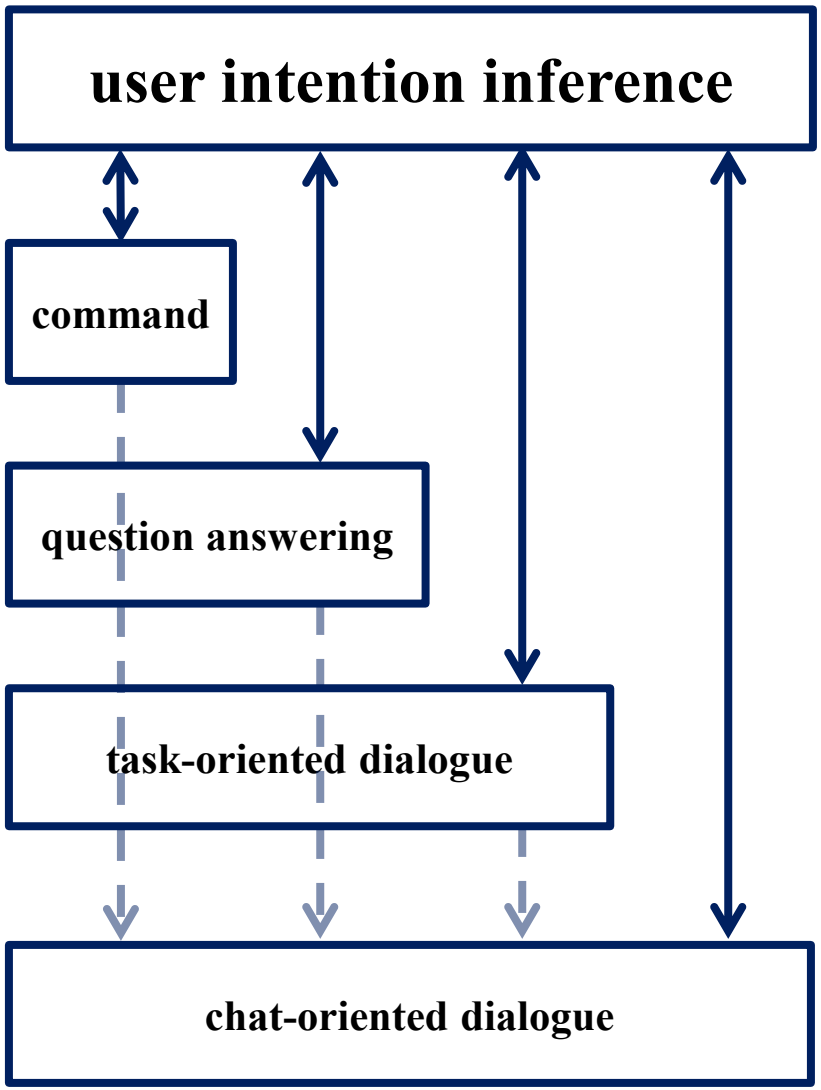
Multiple Dimensions of Complexity



* R. E. Banchs, R. Jiang, S. Kim, A. Niswar, K. H. Yeo (2013), Dialogue Orchestration: integrating different human-computer interaction tasks into a single intelligent conversational agent, White Paper

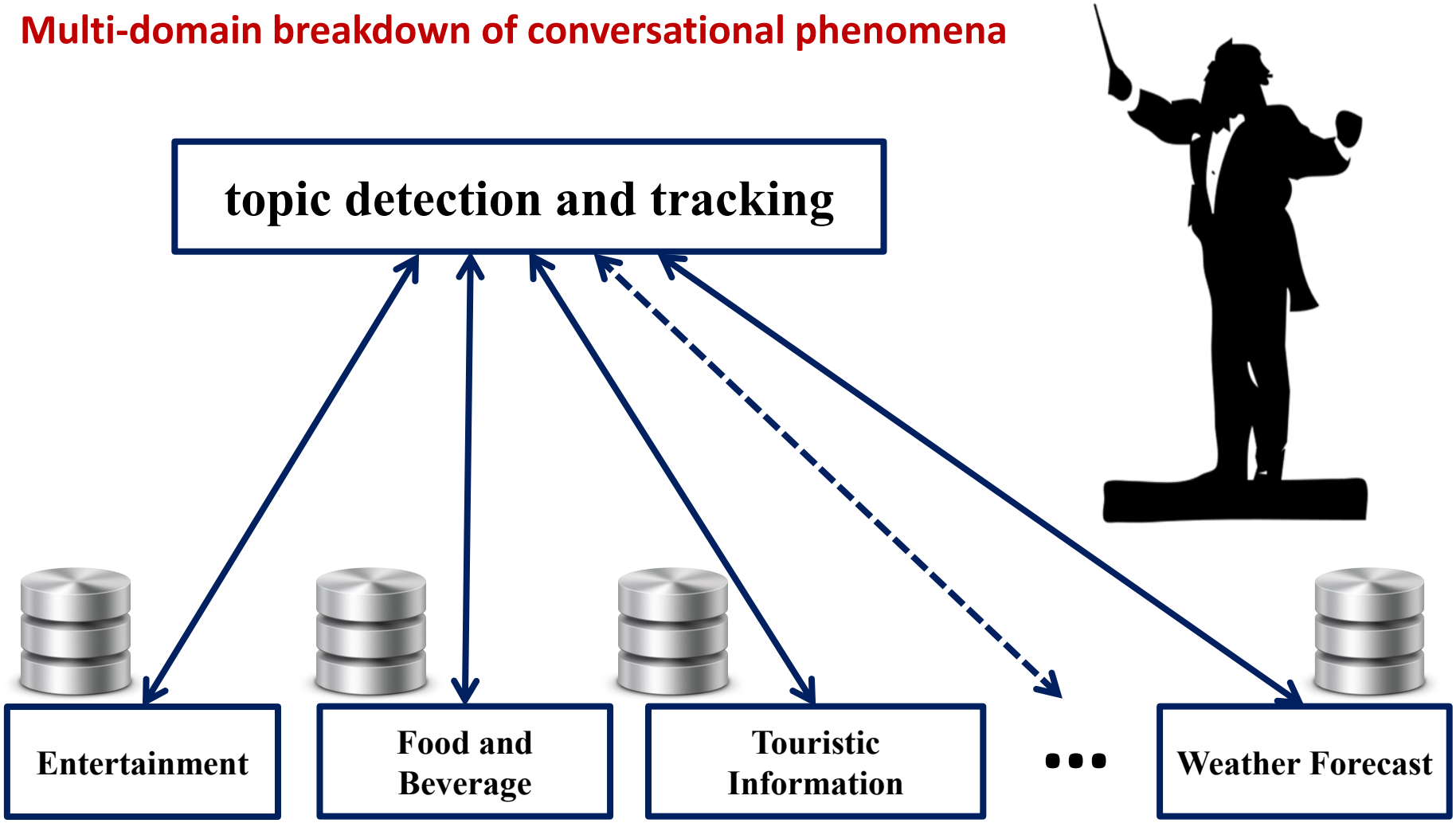
Multiple Levels of Interaction

A hierarchical approach to conversational phenomena



Multiplicity of Domains

Multi-domain breakdown of conversational phenomena



* I. Lee, S. Kim, K. Kim, D. Lee, J. Choi, S. Ryu, and G. G. Lee, "A two-step approach for efficient domain selection in muti-domain dialog systems", in Int'l Workshop on Spoken Dialog Systems, 2012

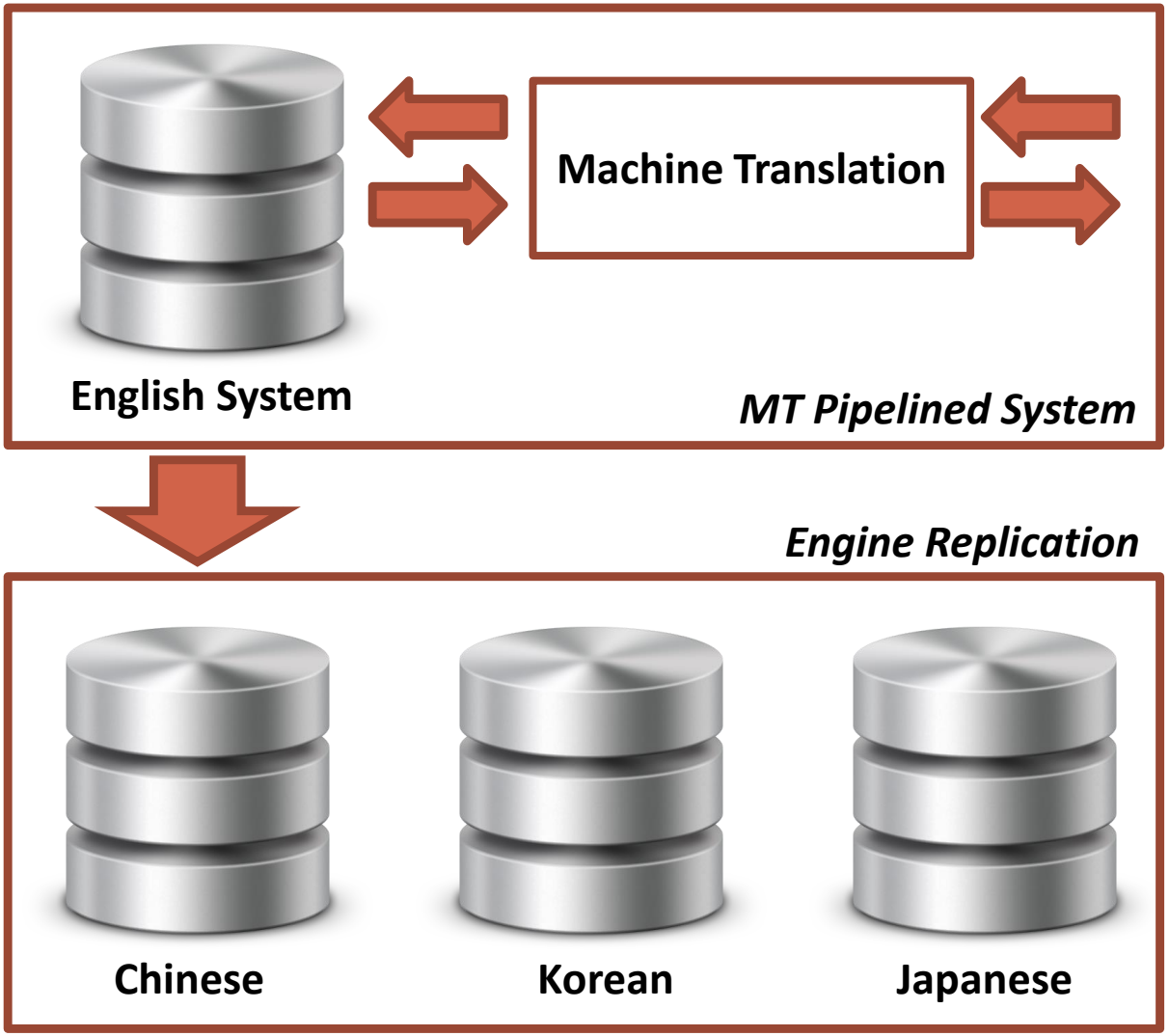
MULTIMODALITY

speech, text, touch screen, image, video, gestures, emotions, sound localization, voice biometrics, face recognition…

Multiplicity of Users

- Face detection
- Sound localization
- Who is talking?
- User profiling
- Thread detection and tracking
- Turn taking

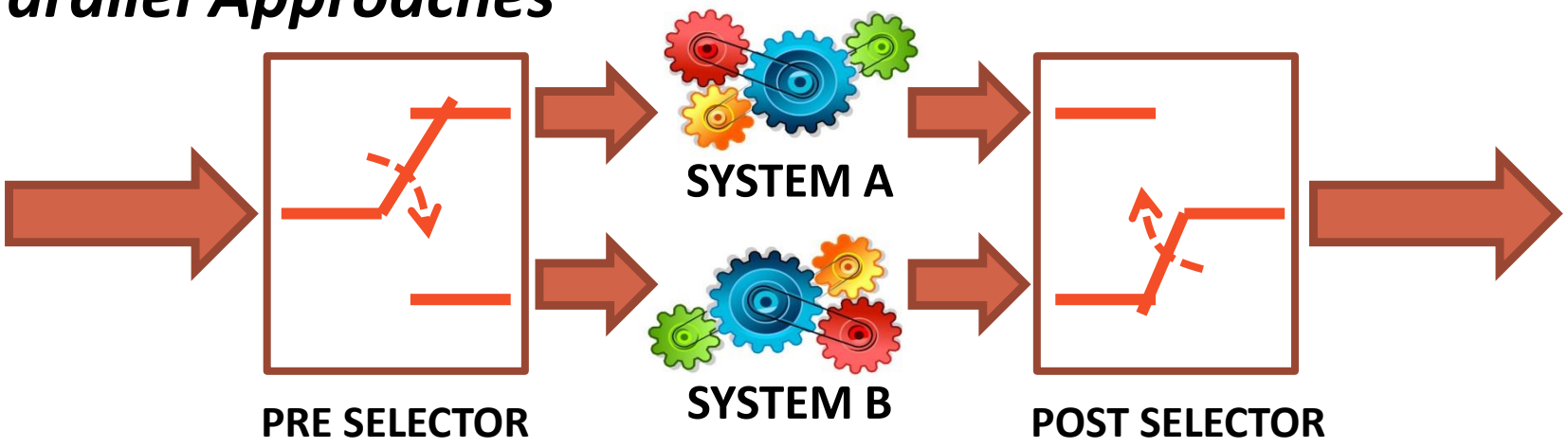
Multiplicity of Languages



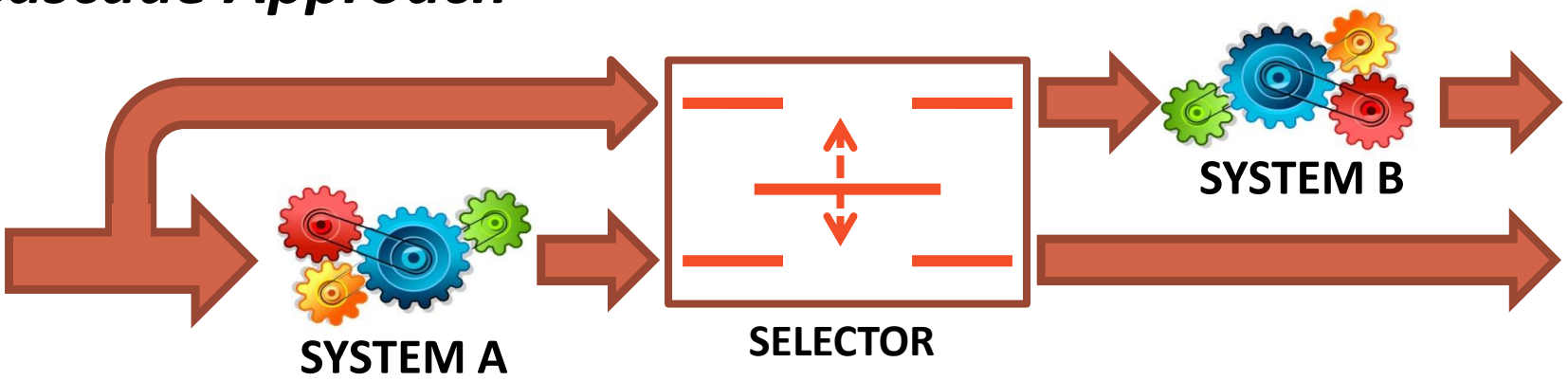
Language Interpreter

Multiplicity of Systems (System Combination)

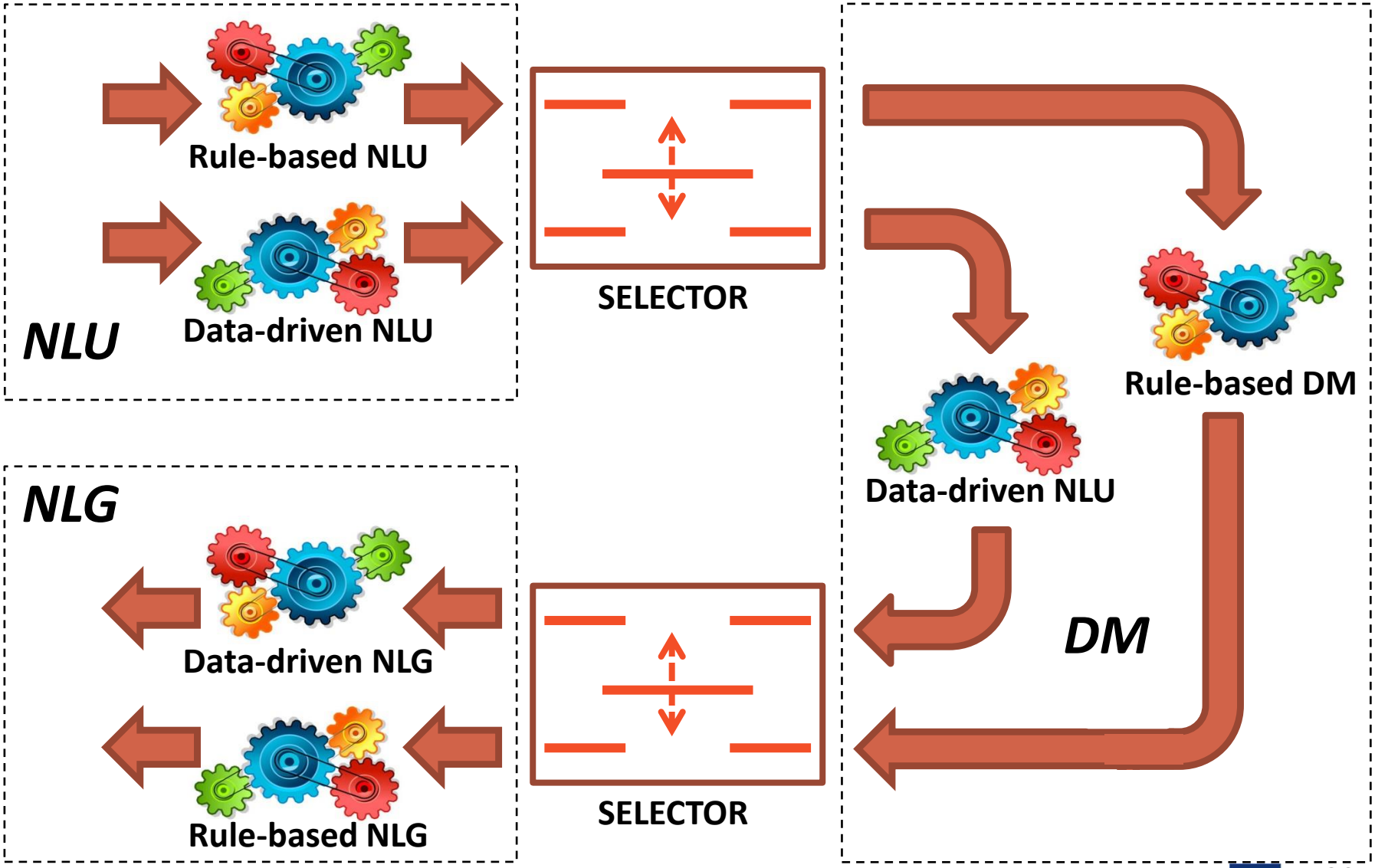
Parallel Approaches



Cascade Approach



Multiplicity of Systems (Hybrid Approaches)



In Summary

- Human-Robot Interaction is complex:
  - Multidimensional Problem
    - *Vision, Language, Tactile, Localization, Navigation, Manipulation, Social Skills*
  - Multidisciplinary Problem
    - *Computer Science, Social Sciences, Psychology, Engineering, Human Factors, User Experience Design*
  - Ill-defined (i.e. non-natural problem)
    - *Human-Human vs. Human-Robot Interactions*
    - *Natural Language vs. Structured Language*
    - *Limited performance of the involved technologies!*

Main References

- D. Spiliotopoulos, I. Androutsopoulos, C. Spyropoulos "Human-robot interaction based on spoken natural language dialogue" *Eur. Workshop Service and Humanoid Robots (ServiceRob)* pp. 1057-1060. http://www.aueb.gr/users/ion/docs/servicerob_paper.pdf
- Austin, J. L. 1962. “How to do things with words”. London: Oxford University Press
- R. E. Banchs, R. Jiang, S. Kim, A. Niswar, K. H. Yeo (2013), Dialogue Orchestration: integrating different human-computer interaction tasks into a single intelligent conversational agent, White Paper
- I. Lee, S. Kim, K. Kim, D. Lee, J. Choi, S. Ryu, and G. G. Lee, “A two-step approach for efficient domain selection in muti-domain dialog systems”, in Int’l Workshop on Spoken Dialog Systems, 2012

Additional References

- Cuayahuitl, Heriberto and Komatani, Kazunori and Skantze, Gabriel (2015) Introduction for speech and language for interactive robots. *Computer Speech & Language*, 34 (1). pp. 83-86. ISSN 0885-2308
- Nikolaos Mavridis, A review of verbal and non-verbal human–robot interactive communication, *Robotics and Autonomous Systems*, Volume 63, Part 1, January 2015, Pages 22–35, http://dx.doi.org/10.1016/j.robot.2014.09.031
- B. S. Lin, H. M. Wang, and L. S. Lee (2001) "A distributed agent architecture for intelligent multi-domain spoken dialogue systems", *IEICE Trans. On Information and Systems*, E84-D(9), pp. 1217–1230.
- The Future of Human-Robot Spoken Dialogue: from Information Services to Virtual Assistants (2015), NII Shonan Meeting, *NII Shonan Meeting Report* (ISSN 2186-7437):No.2015-7, accessed online at http://shonan.nii.ac.jp/shonan/blog/2013/12/10/the-future-of-human-robot-spoken-dialogue-from-information-services-to-virtual-assistants/

RESOURCES

- **KANTRA** - A Natural Language Interface for Intelligent Robots, Thomas Laengle, Tim C. Lueth, Eva Stopp, Gerd Herzog, Gjertrud Kamstrup, http://www.dfki.de/~flint/papers/b114.pdf
- **ROSPEEX**: Speech Communication Toolkit for Robots http://rospeex.org/top/
- **Virtual Human Toolkit**, a collection of modules, tools, and libraries designed to aid and support researchers and developers with the creation of virtual human conversational characters, available online at https://vhtoolkit.ict.usc.edu/

PART 2: SEMANTICS AND PRAGMATICS

Rafael E. Banchs, **Seokhwan Kim**, Luis Fernando D'Haro, Andreea I. Niculescu

Human Language Technology (I^2R, A*STAR)

HAI 2016
4th International Conference on Human-Agent Interaction
4 - 7 OCTOBER
SINGAPORE

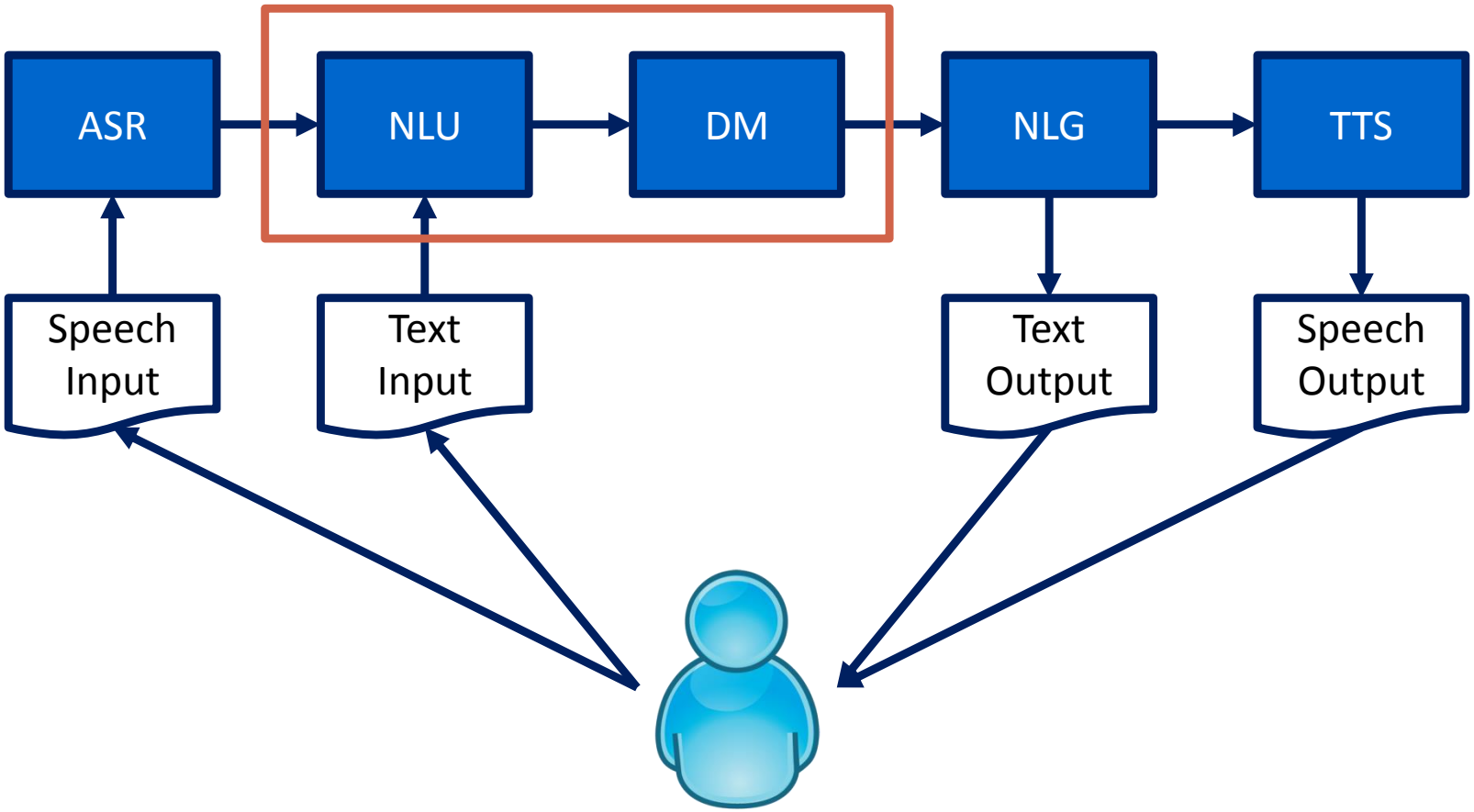
A*STAR Institute for Infocomm Research

Tutorial Content Overview

1. Natural Language in Human-Robot Interaction
   - *Human-Robot Interaction*
   - *The Role of Natural Language*
2. Semantics and Pragmatics
   - *Natural Language Understanding*
   - *Dialogue Management*
3. System Components and Architectures
   - *Frontend System Components (Interfaces)*
   - *Backend System Components*
4. User Experience (UX) Design and Evaluation
   - *UX Design for Speech Interactions*
   - *User Studies and Evaluations*

INTRODUCTION

- Architecture of Conversational Interfaces



Part 2:
Semantics and Pragmatics

NATURAL LANGUAGE UNDERSTANDING

A*STAR
Institute for Infocomm Research

NLU: Introduction

- Natural Language Understanding (NLU)
  - Input
    - Text Input
    - ASR Results for speech inputs
  - Output
    - Semantic Representation

NLU: Introduction

- Semantic Representations for NLU
  - Examples

Show me flights from **Singapore** to **New York** **next Monday**.

Domain	**Flight**
Intent	**Show_flight**
Departure City	**Singapore**
Arrival City	**New York**
Departure Date	**10/10/2016**

NLU: Introduction

- Semantic Representations for NLU
  - Examples

Show me flights from **Singapore** to **New York** **next Monday**.

Domain	**Flight**
Intent	**Show_flight**
Departure City	**Singapore**
Arrival City	**New York**
Departure Date	**10/10/2016**

How can I get to **Orchard Road** from **Fusionopolis** by **MRT**?

Domain	**Transportation**
Intent	**Ask_direction**
Origin	**Fusionopolis**
Destination	**Orchard Road**
Type	**MRT**

NLU: Introduction

- Semantic Representations for NLU
  - Examples

Show me flights from **Singapore** to **New York** **next Monday**.

Domain	**Flight**
Intent	**Show_flight**
Departure City	**Singapore**
Arrival City	**New York**
Departure Date	**10/10/2016**

How can I get to **Orchard Road** from **Fusionopolis** by **MRT**?

Domain	**Transportation**
Intent	**Ask_direction**
Origin	**Fusionopolis**
Destination	**Orchard Road**
Type	**MRT**

I'm looking for a **cheap** **Indian** restaurant in **Orchard**.

Domain	**Restaurant**
Intent	**Request**
Cuisine	**Indian**
Price range	**Cheap**
Neighborhood	**Orchard**

NLU: Introduction

- Semantic Representations for NLU
  - Domain-specific Frame Structure
    - Topic/Domain category
    - User Intent Category
    - Slot/Value Pairs
  - Subtasks of NLU
    - Sentence-level
      - Topic/Domain classification
      - User Intent Identification
    - Word-level
      - Slot-filling

NLU: Methods

- Knowledge-based Approaches
  - Traditional Systems
    - CMU Phoenix: [Ward and Issar 1994]
    - MIT TINA: [Seneff 1992]
    - SRI Gemini: [Dowding et al 1994]
  - Based on Human Knowledge
    - Dictionaries
    - Context Patterns
    - Grammars
    - Regular Expressions

NLU: METHODS

- Knowledge-based Approaches
  - Dictionary Matching

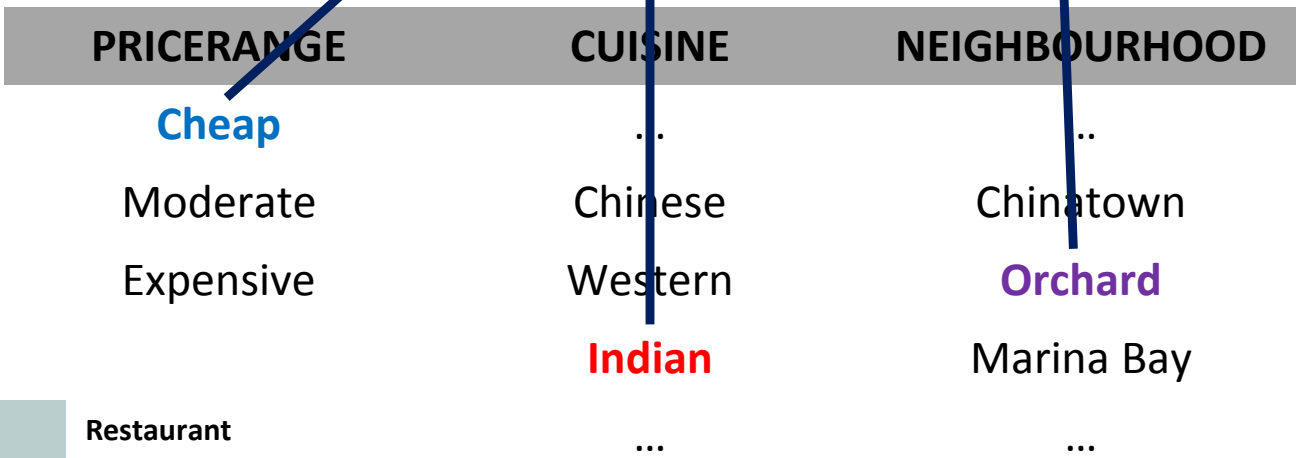
I'm looking for a cheap Indian restaurant in Orchard.

PRICERANGE	CUISINE	NEIGHBOURHOOD
Cheap	…	…
Moderate	Chinese	Chinatown
Expensive	Western	Orchard
	Indian	Marina Bay
	…	…

NLU: METHODS

- Knowledge-based Approaches
  - Dictionary Matching

I'm looking for a **cheap** **Indian** restaurant in **Orchard**.



Domain	**Restaurant**
Intent	**Request**
Cuisine	**Indian**
Price range	**Cheap**
Neighborhood	**Orchard**

NLU: METHODS

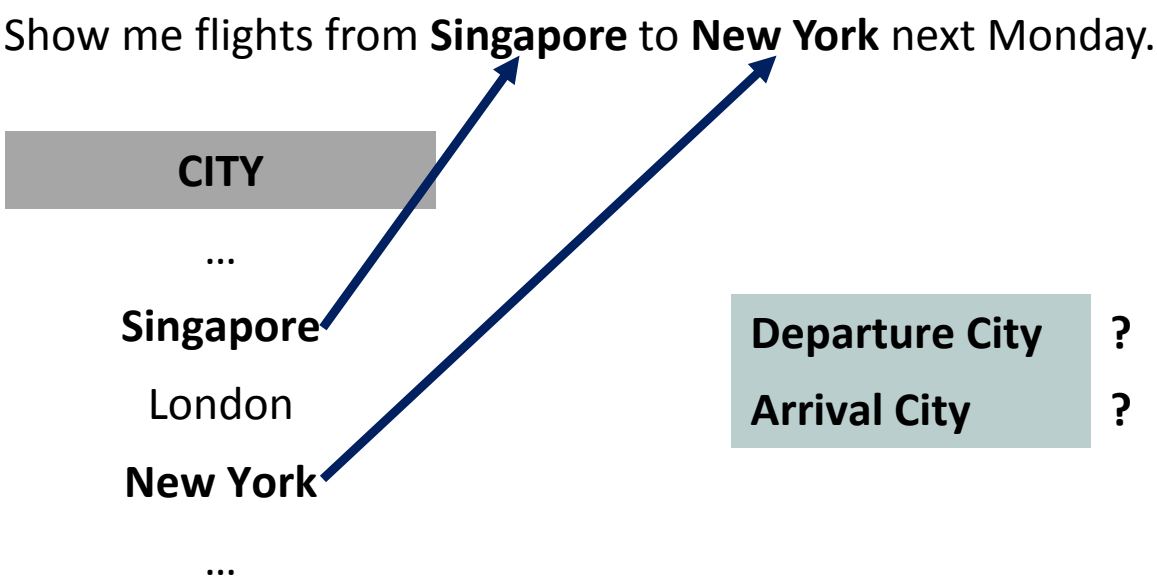
- Knowledge-based Approaches
  - Dictionary Matching

Show me flights from Singapore to New York next Monday.

CITY
…
Singapore
London
New York
…

NLU: METHODS

- Knowledge-based Approaches
  - Dictionary Matching



NLU: METHODS

- Knowledge-based Approaches
  - ❑ Context Pattern/Grammar Matching

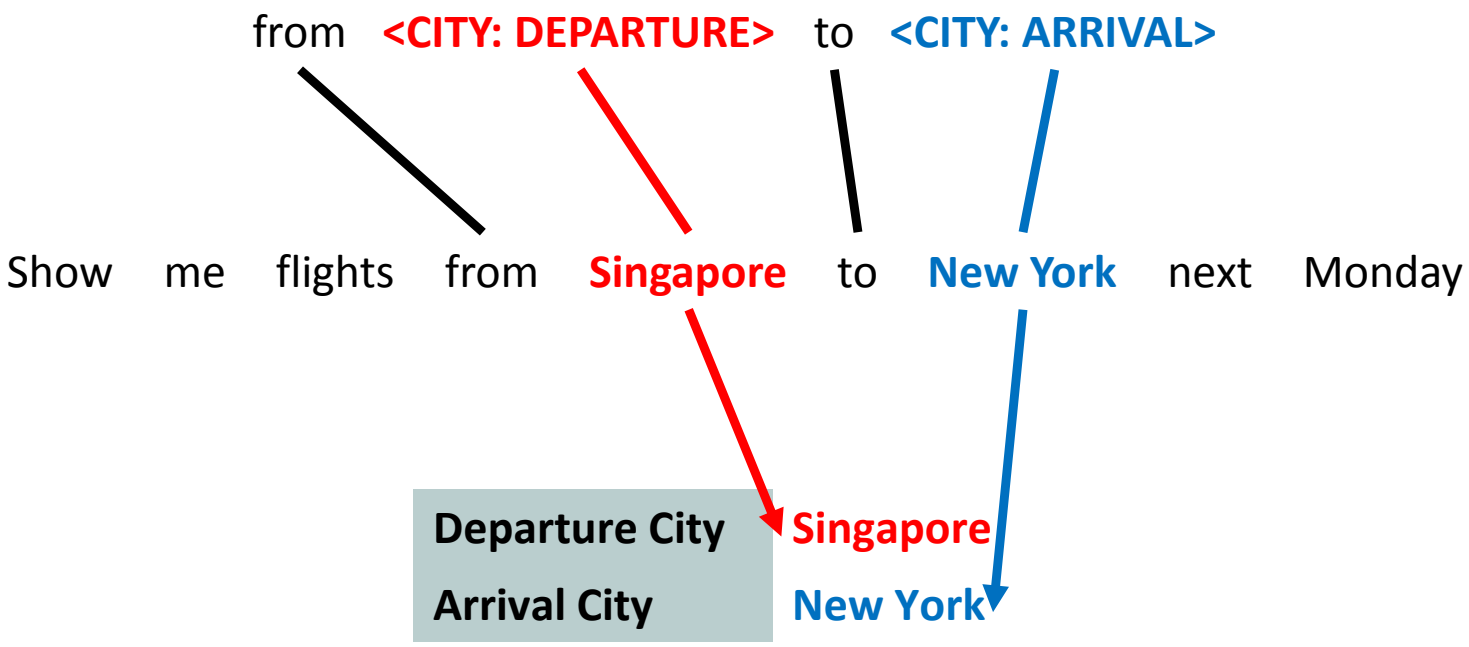
from **<CITY: DEPARTURE>** to **<CITY: ARRIVAL>**

Show me flights from **Singapore** to **New York** next Monday

Departure City
Arrival City

NLU: METHODS

- Knowledge-based Approaches
  - Context Pattern/Grammar Matching



NLU: METHODS

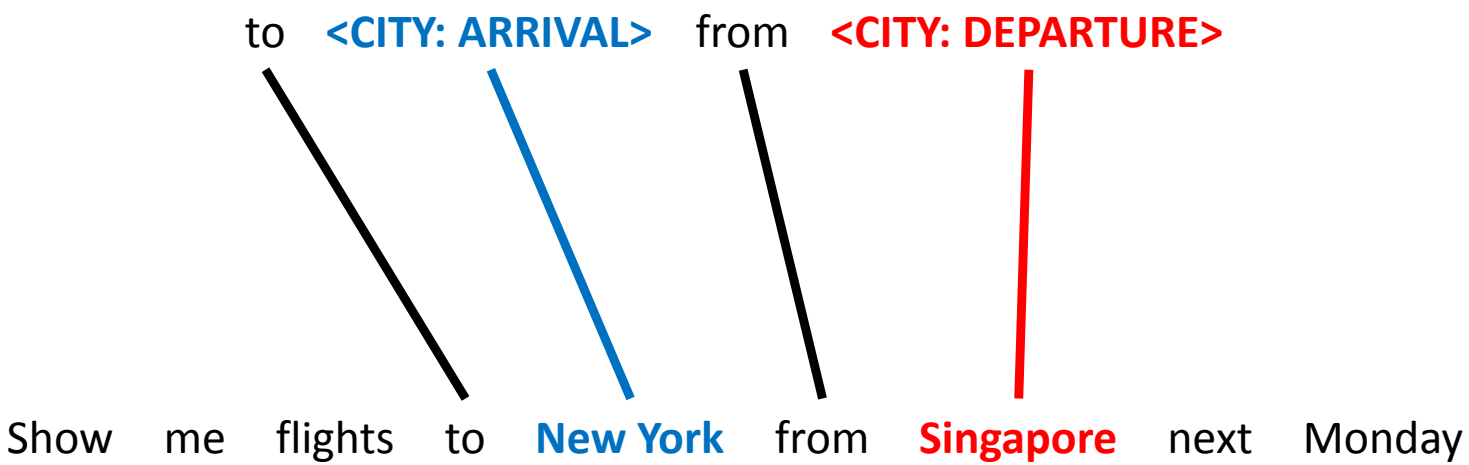
- Knowledge-based Approaches
  - Context Pattern/Grammar Matching

from **<CITY: DEPARTURE>** to **<CITY: ARRIVAL>**

Show me flights to **New York** from **Singapore** next Monday

NLU: METHODS

- Knowledge-based Approaches
  - Context Pattern/Grammar Matching



NLU: Methods

- Knowledge-based Approaches
  - Context Pattern/Grammar Matching
    - Not Scalable
      - Variations

Show me flights heading to **New York** departing from **Singapore**.
I'm looking for flights for **New York** originating in **Singapore**.
Is there any flights leaving **Singapore** for **New York**?
…

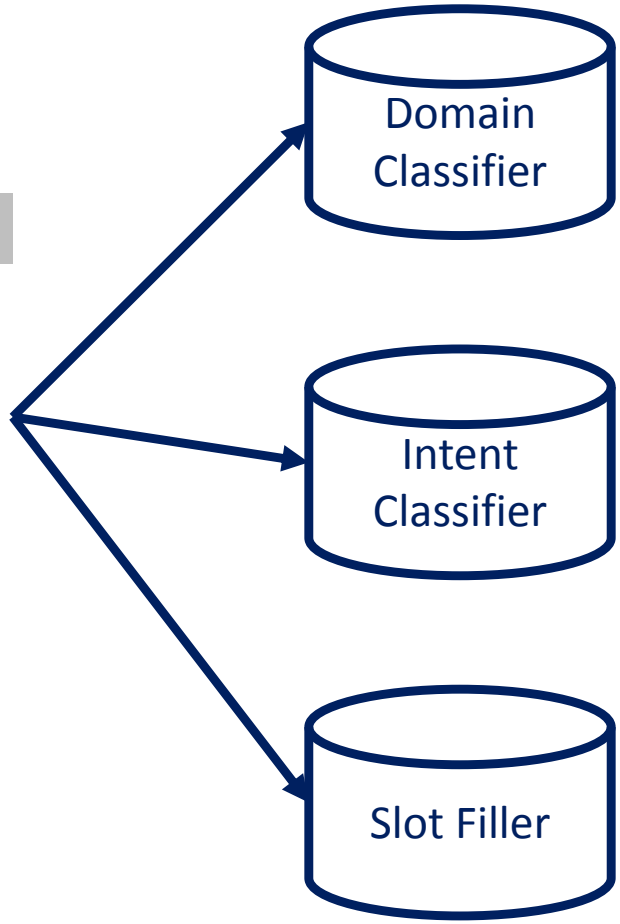
      - Noisy Inputs
        - Typos
        - ASR Errors

NLU: METHODS

- Statistical Approaches

LABELLED CORPUS

…
Show me flights from **Singapore** to **New York**.
Show me flights to **New York** from **Singapore**.
Show me flights heading to **New York** departing from **Singapore**.
I'm looking for flights for **New York** originating in **Singapore**.
Is there any flights leaving **Singapore** for **New York**?
…



NLU: METHODS

- Statistical Approaches
  - Domain/Intent Classification
    - Sentence Classification
    - Training Data: $D = \{(s_1, c_1), \ldots, (s_n, c_n)\}$
      - s_i: the i-th sentence in D
      - $c_i \in C$: manually labelled domain/intent class for s_i
    - Goal: $\text{argmax}_{c \in C} P(c|s)$

INPUT	Show me flights from Singapore to New York next Monday.
DOMAIN	
INTENT	

NLU: METHODS

- Statistical Approaches
  - Slot Filling
    - Sequence Labelling
    - Training Data: $s = \{(w_1, c_1), \ldots, (w_n, c_n)\}$
      - w_i: the i-th word in sentence s
      - $c_i \in C$: manually labelled semantic tag class for w_i
    - Goal: $\text{argmax}_{c \in C} P(c|w)$

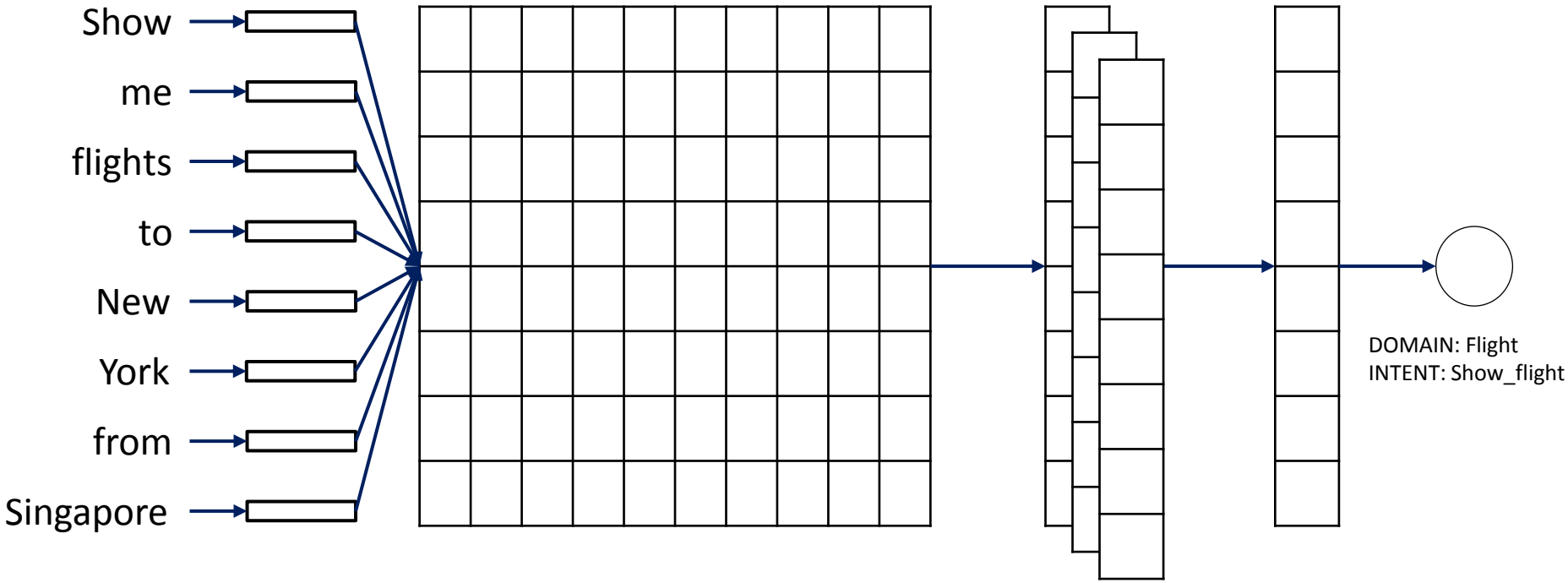
Show	me	flights	from	Singapore	to	New	York	next	Monday
O	O	O	O	$\text{B-CITY}_{\text{DEPARTURE}}$	O	$\text{B-CITY}_{\text{ARRIVAL}}$	$\text{I-CITY}_{\text{ARRIVAL}}$	B-DATE	I-DATE

NLU: Methods

- Statistical Approaches
  - Machine Learning Models
    - Generative Models
      - [Levin and Pieraccini 1995, Miller et al. 1994, He and Young 2005]
    - Discriminative Models
      - [Kuhn and De Mori 1995, Jeong and Lee 2006, Wang and Acero 2006, Raymond and Riccardi 2007, Moschitti et al. 2007, Henderson et al. 2012]
  - Features
    - Bag-of-words
    - Word n-grams
    - Linguistic Pre-processing
    - ASR Hypotheses
    - Confusion Network

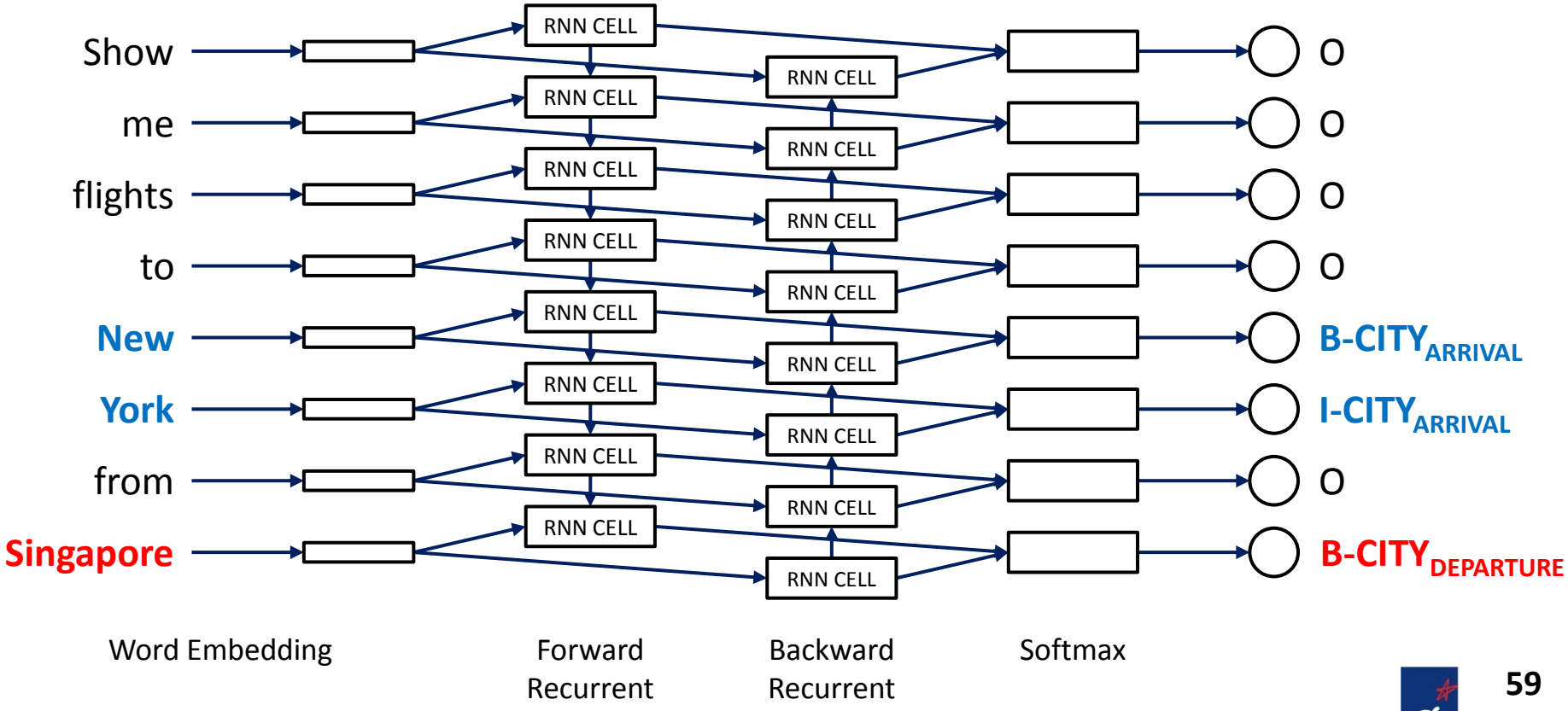
NLU: Recent Trends

- Deep Learning for NLU
  - Convolutional Neural Networks (CNN)
    - [Xu and Sarikaya 2013, Kim et al. 2015]



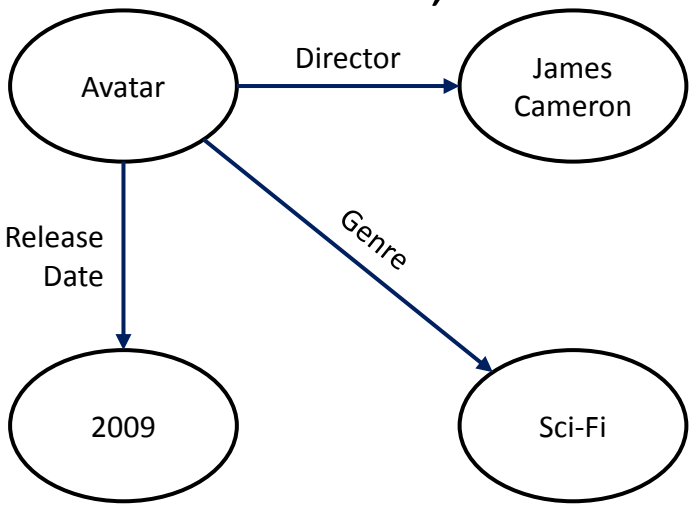
NLU: Recent Trends

- Deep Learning for NLU
  - Recurrent Neural Networks (RNN)
    - [Yao et al. 2013, Mesnil et al. 2013, Yao et al. 2014]



NLU: Recent Trends

- Leveraging External Knowledge for NLU
  - Knowledge Graph
    - [Heck and Hakkani-Tür 2012, Heck et al. 2013]



	Intent	Slots
Who directed Avatar?	Find_director	Movie
Show movies James Cameron directed	Find_movie	Director
Find me some Sci-Fi movies	Find_movie	Genre
Did James Cameron direct Avatar?	Find_movie/director	Director/Movie

NLU: Recent Trends

- Leveraging External Knowledge for NLU
  - Wikipedia [Kim et al. 2014, Kim et al. 2015]

Avatar (2009 film)

From Wikipedia, the free encyclopedia

Avatar (marketed as ***James Cameron's Avatar***) is a 2009 American[7][8] epic science fiction film directed, written, produced, and co-edited by James Cameron, and starring Sam Worthington, Zoe Saldana, Stephen Lang, Michelle Rodriguez, and Sigourney Weaver. The film is set in the mid-22nd century, when humans are colonizing Pandora, a lush habitable moon of a gas giant in the Alpha Centauri star system, in order to mine the mineral unobtanium,[9][10] a room-temperature superconductor.[11] The expansion of the mining colony threatens the continued existence of a local tribe of Na'vi – a humanoid species indigenous to Pandora. The film's title refers to a genetically engineered Na'vi body with the mind of a remotely located human that is used to interact with the natives of Pandora.[12]

Development of *Avatar* began in 1994, when Cameron wrote an 80-page treatment for the film.[13][14] Filming was supposed to take place after the completion of Cameron's 1997 film *Titanic*, for a planned release in 1999,[15] but according to Cameron, the necessary technology was not yet available to achieve his vision of the film.[16] Work on the language of the film's extraterrestrial beings began in 2005, and Cameron began developing the screenplay and fictional universe in early 2006.[17][18] *Avatar* was officially budgeted at $237 million.[3] Other estimates put the cost between $280 million and $310 million for production and at $150 million for promotion.[19][20][21] The film made extensive use of new motion capture filming techniques,[22] and was released for traditional viewing, 3D viewing (using the RealD 3D, Dolby 3D, XpanD 3D, and IMAX 3D formats), and for "4D" experiences in select South Korean theaters.[23] The stereoscopic filmmaking was touted as a breakthrough in cinematic technology.[24]

Text

Hyperlinks

Infobox

Categories

Avatar

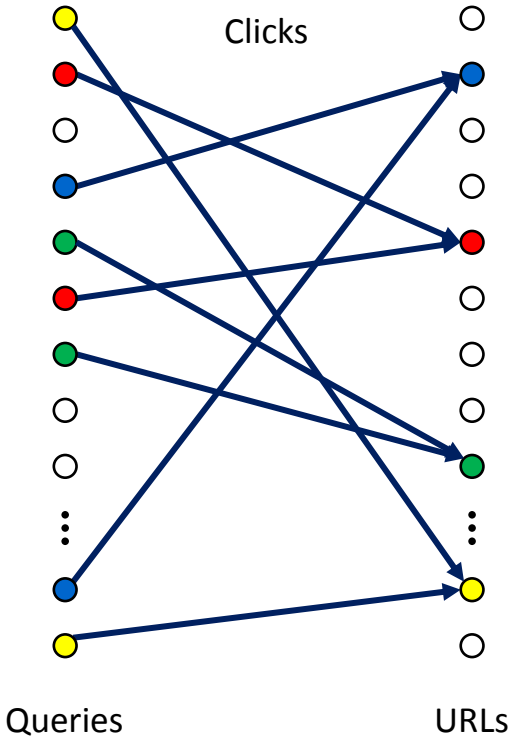
Theatrical release poster

Directed by	James Cameron
Produced by	James Cameron Jon Landau
Written by	James Cameron
Starring	Sam Worthington Zoe Saldana Stephen Lang Michelle Rodriguez Sigourney Weaver
Music by	James Horner
Cinematography	Mauro Fiore
Edited by	James Cameron John Refoua Stephen E. Rivkin
Production company	Lightstorm Entertainment Dune Entertainment Ingenious Film Partners
Distributed by	20th Century Fox
Release dates	December 10, 2009 (London premiere) December 17, 2009 (United Kingdom) December 18, 2009

Categories: 2009 films | English-language films | Avatar (2009 film) | 2000s 3D films | 2000s action films | 2000s science fiction adventure films | Science fiction adventure films | 2000s adventure films | 20th Century Fox films | 22nd century in fiction | American epic films | American films | American science fiction action films | Artificial uterus in fiction | Best Film Empire Award winners | Dune Entertainment films | Environmental films | Fictional-language films | Film scores by James Horner | Films about cloning | Films about consciousness | Films about extraterrestrial life | Films about paraplegics or quadriplegics | Films about rebellions | Films about twins | Films about technology | Films about telepresence | Films directed by James Cameron | Films set in the 22nd century | Films set on fictional moons | Films shot in California | Films shot in Hawaii | Films shot in New Zealand | Films that won the Best Visual Effects Academy Award | Films using computer-generated imagery | Films whose art director won the Best Art Direction Academy Award | Films whose cinematographer won the Best Cinematography Academy Award | Holography in films | IMAX films | Lightstorm Entertainment films | Military science fiction films | Performance capture in film | Planetary romances | Rebellions in fiction | Rotoscoped films | Science fiction war films | Screenplays by James Cameron | Social science fiction films | Space adventure films | Transhumanism in film | 2000s science fiction films

NLU: Recent Trends

- Leveraging External Knowledge for NLU
  - Query Click Logs
    - [Tür et al. 2011]



Query	URL
weather in Singapore	https://weather.com/...
zika symptom	http://www.who.int/...
where to eat Chilli Crab	http://www.hungrygowhere.com/...
MRT operating hours	http://www.smrt.com.sg/...
flight from singapore to new york	https://www.expedia.com/...
singapore flyer ticket cost	www.singaporeflyer.com/...
mbs contact	www.marinabaysands.com/...

NLU: References

- Y. Wang, L. Deng, and A. Acero. September 2005, Spoken Language Understanding: An introduction to the statistical framework. IEEE Signal Processing Magazine, 27(5)
- W. Ward and S. Issar. "Recent improvements in the CMU spoken language understanding system." Proceedings of the workshop on Human Language Technology. Association for Computational Linguistics, 1994.
- S. Seneff. "TINA: A natural language system for spoken language applications." *Computational linguistics* 18.1 (1992): 61-86.
- J. Dowding, R. Moore, F. Andry, and D. Moran, “Interleaving syntax and semantics in an efficient bottom-up parser,” in Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Maxico, June 1994, pp. 110–116.
- E. Levin and R. Pieraccini, “CHRONUS, The next generation,” in Proc. 1995 ARPA Spoken Language Systems, Technology Workshop, Austin, TX, Jan. 1995.
- S. Miller, R. Bobrow, R Schwartz, and R. Ingria, “Statistical language processing using hidden understanding models,” in Proc. of 1994 ARPA Spoken Language Systems, Technology Workshop, Princeton, NJ, Mar. 1994.
- Y. He and S. Young. Semantic processing using the hidden vector state model. Computer Speech and Language, 19:85–106. 2005.
- R. Kuhn and R. De Mori, “The application of semantic classi- fication trees to natural language understanding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 449–460, 1995.
- M. Jeong and G. G. Lee, “Exploiting non-local features for spoken language understanding.” in ACL, 2006.
- Y.-Y. Wang and A. Acero, “Discriminative models for spoken language understanding,” in ICSLP, 2006
- C. Raymond and G. Riccardi. "Generative and discriminative algorithms for spoken language understanding." *INTERSPEECH*. 2007.
- A. Moschitti, G. Riccardi, and C. Raymond, “Spoken language understanding with kernels for syntactic/semantic structures,” in ASRU, 2007, pp. 183–188
- M. Henderson, M. Gasic, B. Thomson, P. Tsiakoulis, K. Yu, and S. Young, “Discriminative spoken language understanding using word confusion networks,” in IEEE SLT Workshop, 2012.

NLU: References

- K. Yao, G. Zweig, M. Hwang, Y. Shi, and Dong Yu, “Recurrent neural networks for language understanding,” in INTERSPEECH, 2013.
- G. Mesnil, X. He, L. Deng, and Y. Bengio, “Investigation of recurrent-neural-network architectures and learning methods for language understanding,” in INTERSPEECH, 2013.
- P. Xu and R. Sarikaya, “Convolutional neural network based triangular CRF for joint detection and slot filling,” in ASRU, 2013.
- K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi. Spoken language understanding using long short-term memory neural networks. In Spoken Language Technology Workshop (SLT), 2014 IEEE (pp. 189-194). IEEE. 2014.
- S. Kim, R. E. Banchs, and H. Li. Exploring Convolutional and Recurrent Neural Networks in Sequential Labelling for Dialogue Topic Tracking. ACL 2016.

- L. Heck and D. Hakkani-Tür. "Exploiting the semantic web for unsupervised spoken language understanding." Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, 2012.
- L. Heck, D. Hakkani-Tür, and G. Tür. "Leveraging knowledge graphs for web-scale unsupervised semantic parsing." INTERSPEECH. 2013.
- G. Tür, D. Hakkani-Tür, D. Hillard, and A. Celikyilmaz, Towards Unsupervised Spoken Language Understanding: Exploiting Query Click Logs for Slot Filling. In *INTERSPEECH* (2011).

- S. Kim, R. E. Banchs, and H. Li. A Composite Kernel Approach for Dialog Topic Tracking with Structured Domain Knowledge from Wikipedia, ACL, 2014.
- S. Kim, R. E. Banchs, and H. Li. Wikification of Concept Mentions within Spoken Dialogues Using Domain Constraints from Wikipedia, EMNLP, 2015.

NLU: Resources

- Datasets
  - Airline Travel Information System (ATIS)
    - https://catalog.ldc.upenn.edu/docs/LDC93S4B/corpus.html
  - French MEDIA
    - http://catalog.elra.info/product_info.php?products_id=1057
  - Cambridge
    - http://camdial.org/~mh521/dstc/
  - TourSG
    - http://www.colips.org/workshop/dstc4/data.html
- Toolkits
  - CSLU toolkit
    - http://www.cslu.ogi.edu/toolkit/
  - CMU Phoenix
    - http://wiki.speech.cs.cmu.edu/olympus/index.php/Phoenix
  - Triangular-chain CRFs
    - https://github.com/minwoo/TriCRF
  - Language Understanding Intelligent Service (LUIS)
    - https://www.luis.ai/

TEA BREAK

30 minutes

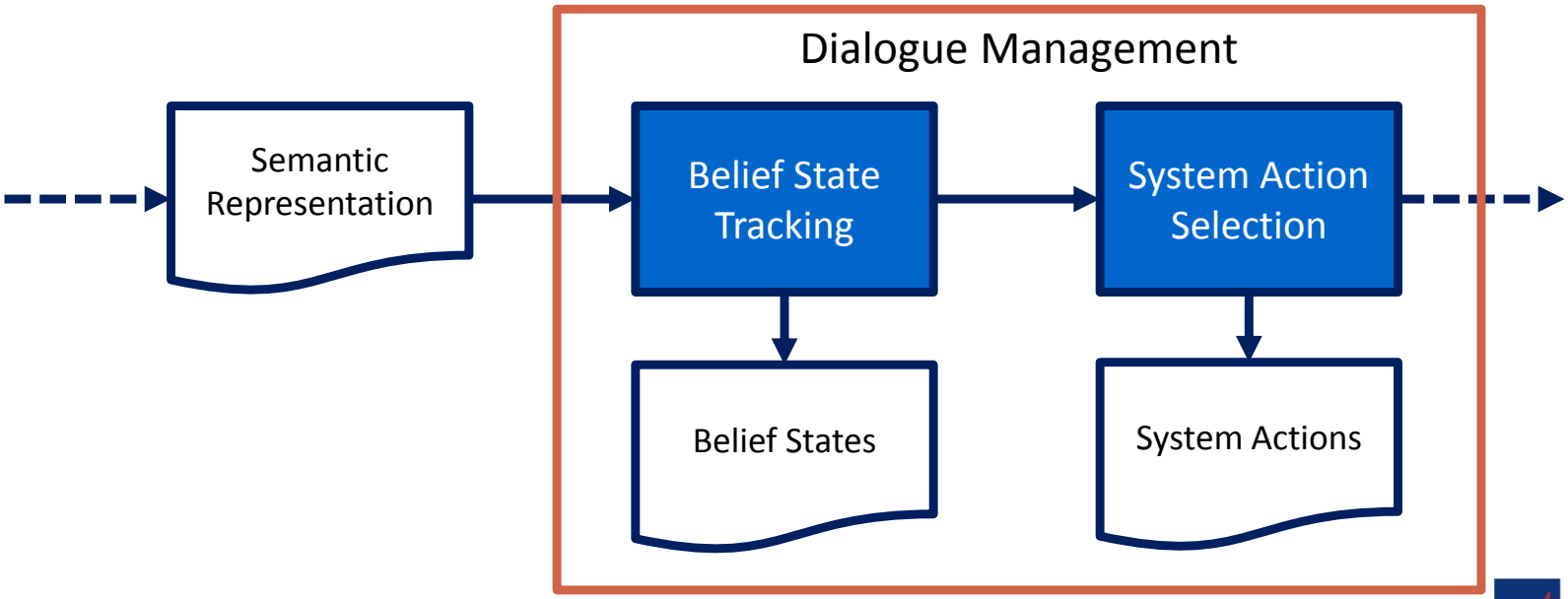
Part 2:
Semantics and Pragmatics

DIALOGUE MANAGEMENT

A*STAR
Institute for Infocomm Research

DM: Introduction

- Dialogue Management (DM)
  - Input: Semantic Representations from NLU
  - Output
    - Belief States
    - System Actions



DM: Introduction

- Dialogue State/Belief Tracking
  - Defines dialogue states representations
    - Distribution over the state hypotheses
  - Updates them at each moment on conversation
    - Considering dialogue history

DM: Introduction

- Dialogue State/Belief Tracking

	Utterance	NLU	Food	
S	Hello, How may I help you?			
U	I need a **Persian** restaurant in the south part of town.	0.2 Inform(food=Persian) 0.8 Inform(area=South)	**0.2**	**Persian**
S	What kind of food would you like?			
U	**Persian**.	0.9 Inform(food=Persian)	**0.8**	**Persian**
S	I'm sorry but there is no restaurant serving persian food			
U	How about **Portuguese** food?	0.7 Inform(food=Portuguese)	0.4 **0.6**	Persian **Portuguese**
S	Are you looking for Portuguese food?			
U	Yes.	1.0 Affirm	0.1 **0.9**	Persian **Portuguese**
S	Nandos is a nice place in the south of town serving tasty Portuguese food.			

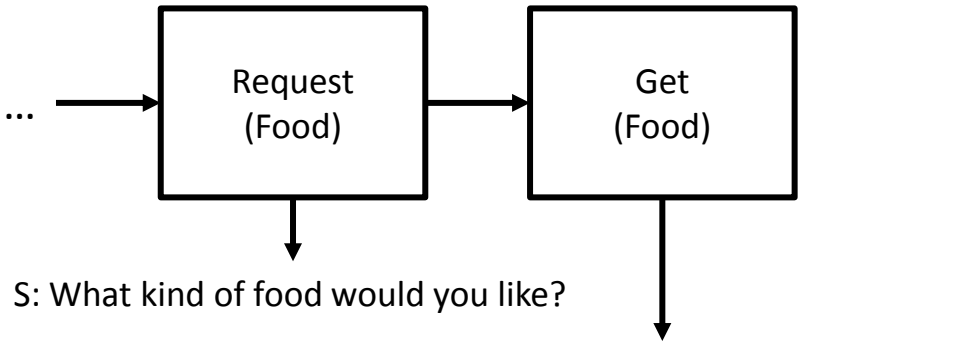
DM: Introduction

- System Action Decision
  - Map from belief state to system action
  - Mapping is called the policy

	Utterance	Belief State		System Action
S	Hello, How may I help you?			
U	I need a **Persian** restaurant in the south part of town.	**0.2**	**Persian**	**Request(food)**
S	What kind of food would you like?			
U	**Persian**.	**0.8**	**Persian**	**Canthelp(food:Persian)**
S	I'm sorry but there is no restaurant serving persian food			
U	How about **Portuguese** food?	0.4	Persian	**Confirm(food:Portuguese)**
S	Are you looking for Portuguese food?	**0.6**	**Portuguese**	
U	Yes.	0.1	Persian	**Offer(place:Nandos)**
S	Nandos is a nice place in the south of town serving tasty Portuguese food.	**0.9**	**Portuguese**	

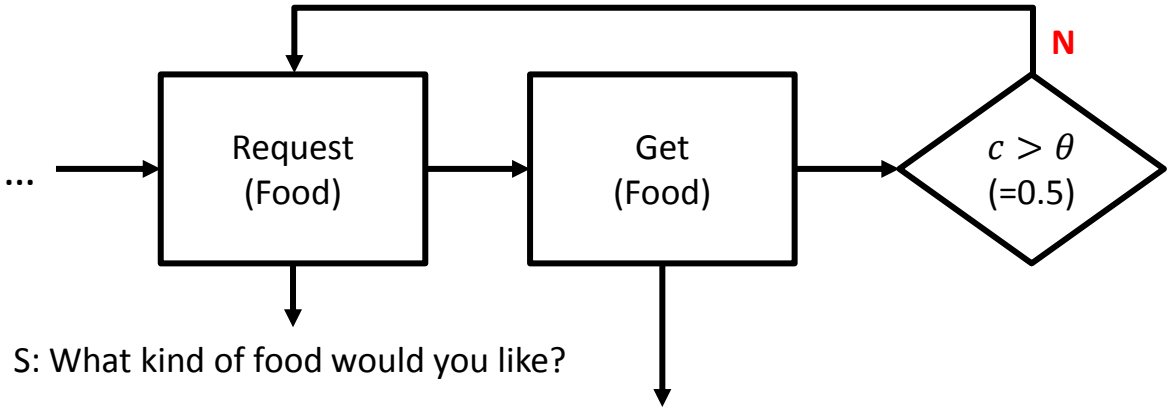
DM: METHODS

- Rule-based Approaches
  - [McTear 1998, Traum and Larsson 2003, Pieraccini and Huerta 2005]



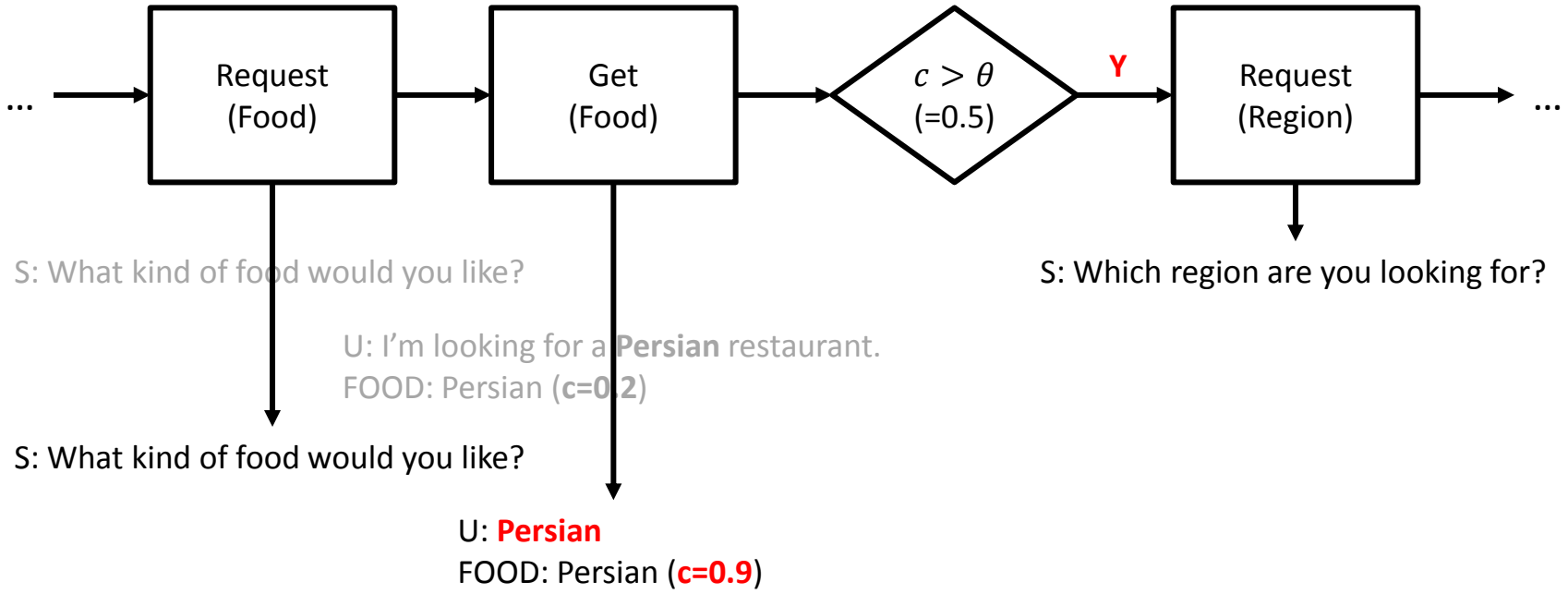
DM: METHODS

- Rule-based Approaches
  - [McTear 1998, Traum and Larsson 2003, Pieraccini and Huerta 2005]



DM: METHODS

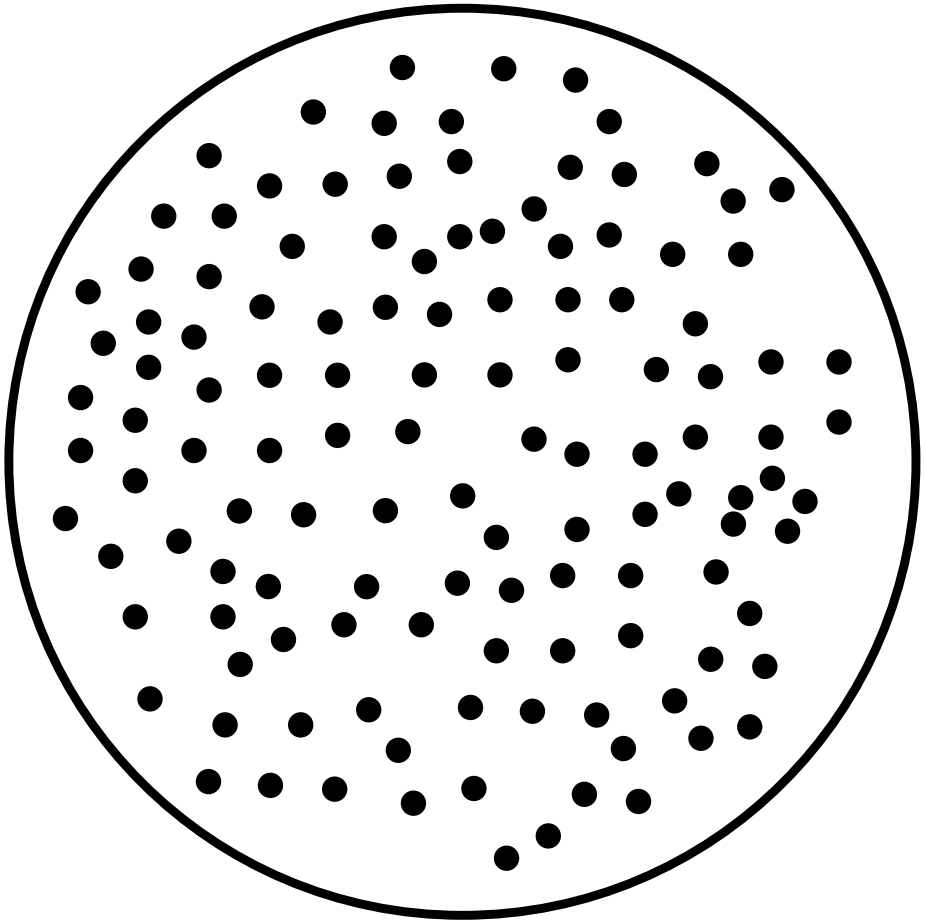
- Rule-based Approaches
  - [McTear 1998, Traum and Larsson 2003, Pieraccini and Huerta 2005]



DM: METHODS

- Example-based Approaches
  - [Lee et al. 2009]

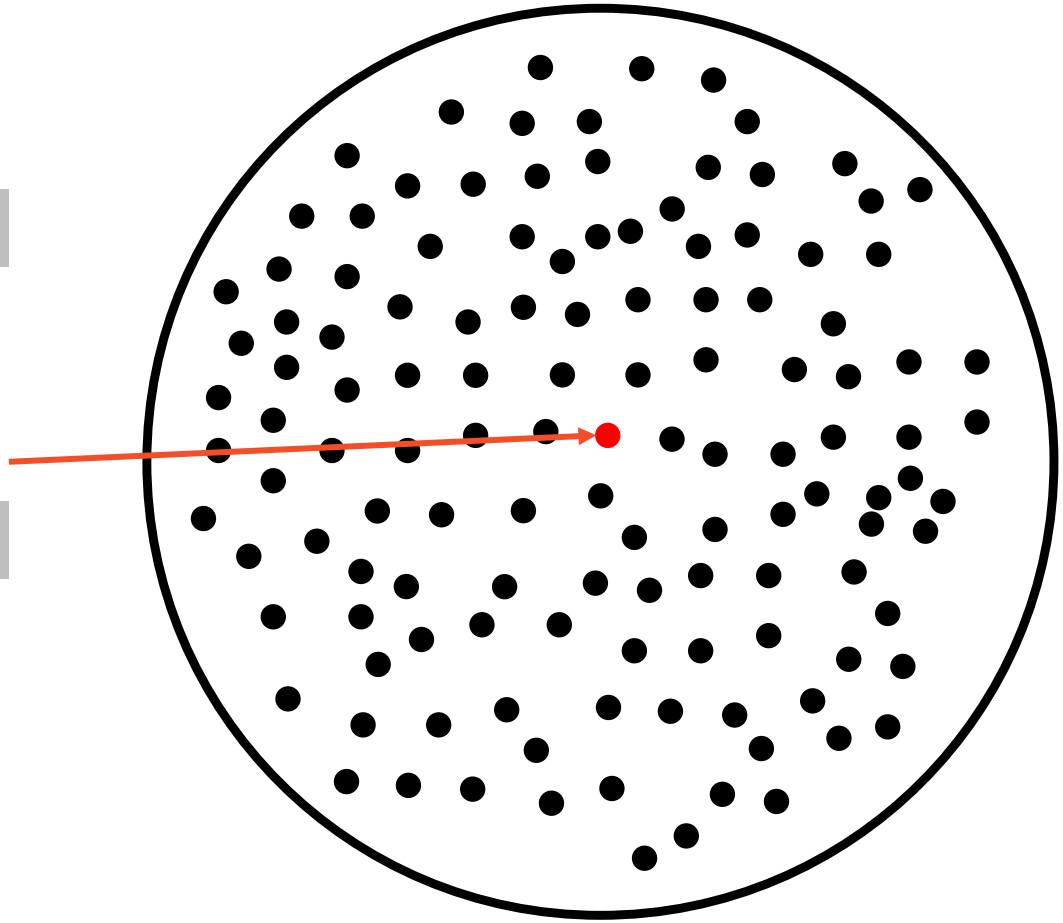
NLU Results	
DOMAIN	RESTAURANT: 0.8
INTENT	STATEMENT: 0.95
SLOT_FOOD	PERSIAN: 0.8
Discourse History Information	
PREVIOUS INTENT	REQUEST: 0.7
FILLED SLOT VECTOR	[0,1,0,1,1,0,0,1,1,1]



DM: METHODS

- Example-based Approaches
  - [Lee et al. 2009]

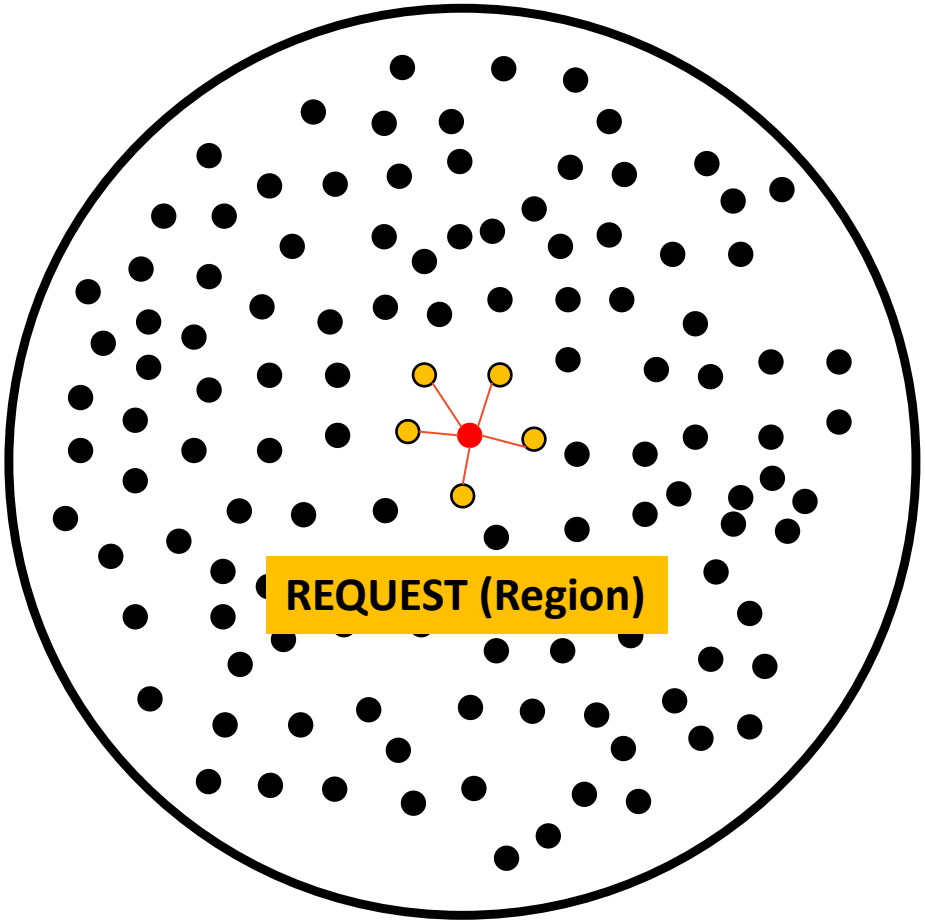
NLU Results	
DOMAIN	RESTAURANT: 0.8
INTENT	STATEMENT: 0.95
SLOT_FOOD	PERSIAN: 0.8
Discourse History Information	
PREVIOUS INTENT	REQUEST: 0.7
FILLED SLOT VECTOR	[0,1,0,1,1,0,0,1,1,1]



DM: METHODS

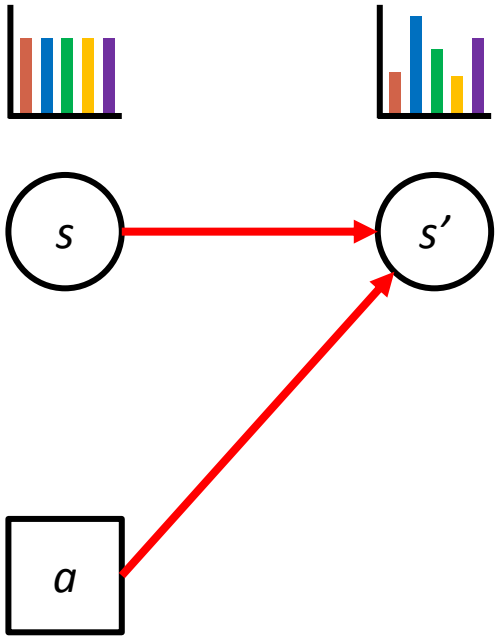
- Example-based Approaches
  - [Lee et al. 2009]

NLU Results	
DOMAIN	RESTAURANT: 0.8
INTENT	STATEMENT: 0.95
SLOT_FOOD	PERSIAN: 0.8
Discourse History Information	
PREVIOUS INTENT	REQUEST: 0.7
FILLED SLOT VECTOR	[0,1,0,1,1,0,0,1,1,1]



DM: METHODS

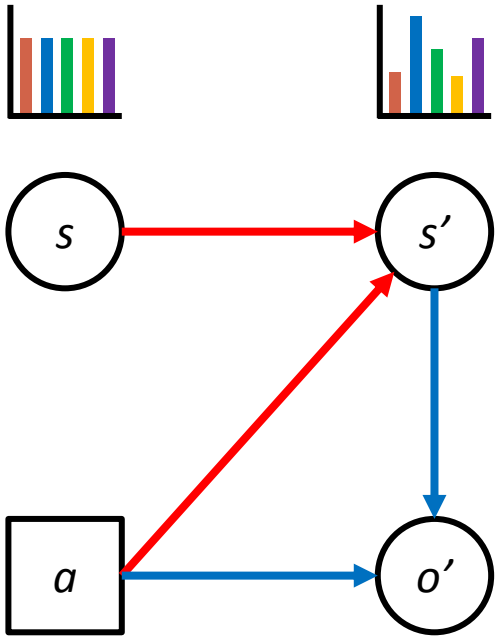
- Statistical Approaches
  - [Williams and Young 2007]



$$P(s'|s,a) = P(g',u',h'|g,u,h,a)$$
$$\approx \underbrace{P(g'|g,a)}_{\text{Goal}} \cdot \underbrace{P(u'|g',h,a)}_{\text{User Action}} \cdot \underbrace{P(h'|u',g',h,a)}_{\text{History}}$$

DM: METHODS

- Statistical Approaches
  - [Williams and Young 2007]



$$P(s'|s,a) = P(g',u',h'|g,u,h,a)$$
$$\approx \boxed{P(g'|g,a)} \cdot \boxed{P(u'|g',h,a)} \cdot \boxed{P(h'|u',g',h,a)}$$

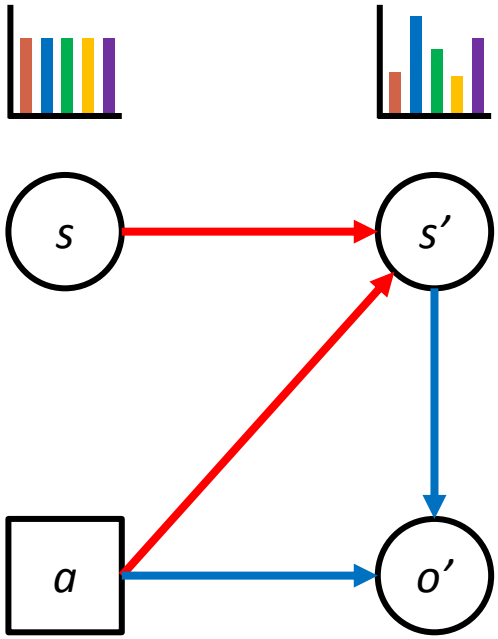
Goal | User Action | History

$$P(o'|s',a) = P(o'|g',u',h,' a) \approx \boxed{P(o'|u',a)}$$

ASR Model

DM: METHODS

- Statistical Approaches
  - [Williams and Young 2007]



$$P(s'|s,a) = P(g',u',h'|g,u,h,a)$$

$$\approx \boxed{P(g'|g,a)} \cdot \boxed{P(u'|g',h,a)} \cdot \boxed{P(h'|u',g',h,a)}$$

Goal | User Action | History

$$P(o'|s',a) = P(o'|g',u',h,' a) \approx \boxed{P(o'|u',a)}$$

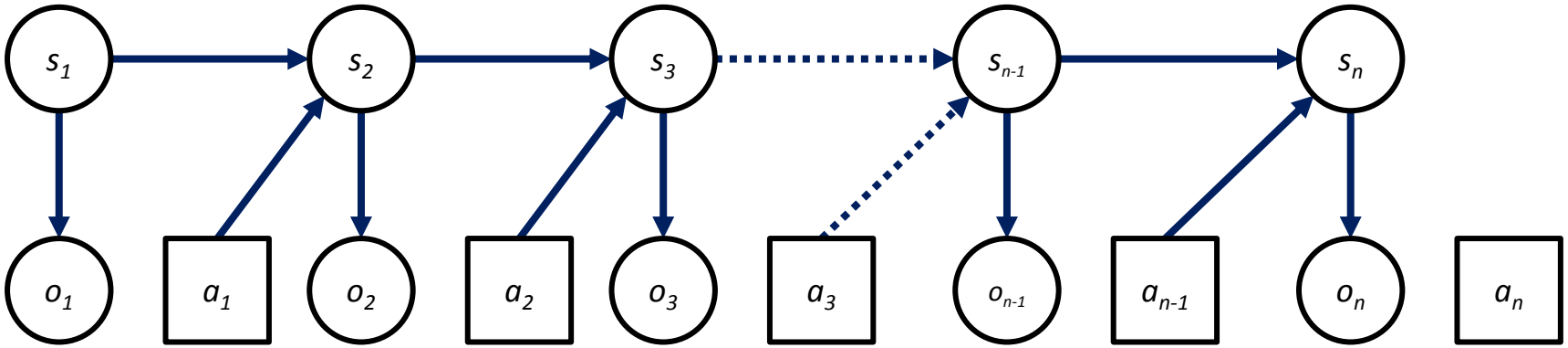
ASR Model

$$\boxed{b'(s')} = \eta \cdot P(o'|s,a) \cdot \sum_s P(s'|s,a) \cdot \boxed{b(s)}$$

Current State | Previous State

DM: METHODS

- Statistical Approaches
  - [Williams and Young 2007]

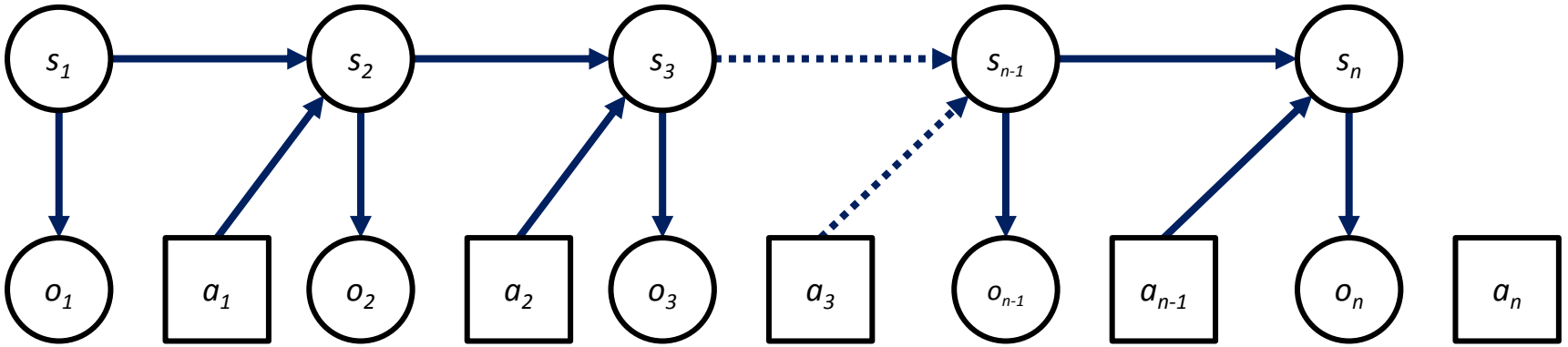


$$\pi(b_1) = a_1 \qquad \pi(b_2) = a_2 \qquad \pi(b_3) = a_3 \qquad \pi(b_{n-1}) = a_{n-1} \qquad \pi(b_n) = a_n$$

Policy

DM: METHODS

- Statistical Approaches
  - [Williams and Young 2007]



$$\pi(b_1) = a_1 \qquad \pi(b_2) = a_2 \qquad \pi(b_3) = a_3 \qquad \pi(b_{n-1}) = a_{n-1} \qquad \pi(b_n) = a_n$$

Policy

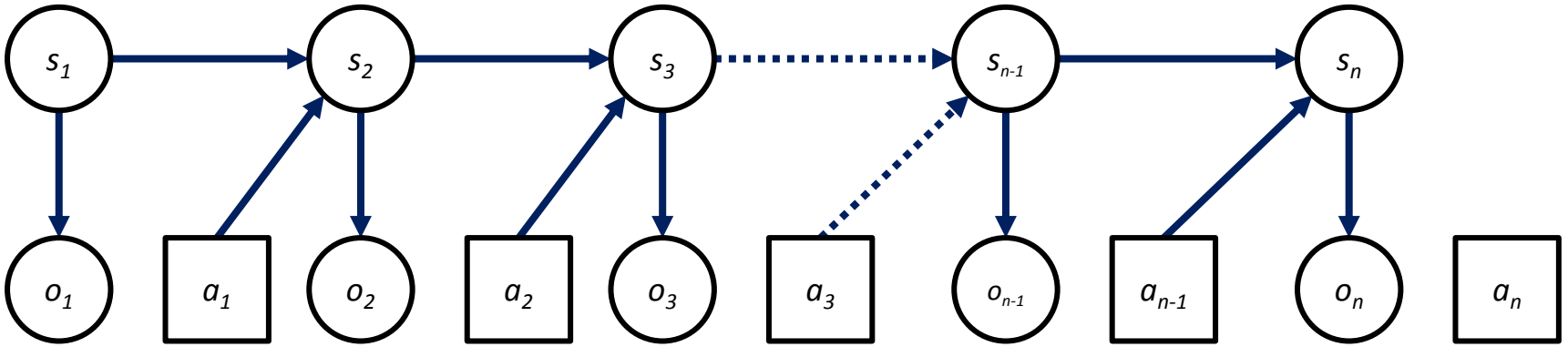
Reward Calculation

Rewards

$$R(b_1, a_1) + R(b_2, a_2) + R(b_3, a_3) + R(b_{n-1}, a_{n-1}) + R(b_n, a_n)$$

DM: METHODS

- Statistical Approaches
  - [Williams and Young 2007]



$$\pi(b_1) = a_1 \qquad \pi(b_2) = a_2 \qquad \pi(b_3) = a_3 \qquad \pi(b_{n-1}) = a_{n-1} \qquad \pi(b_n) = a_n$$

New Policy

Reward Calculation

Reinforcement Learning

Rewards

$$R(b_1, a_1) + R(b_2, a_2) + R(b_3, a_3) + R(b_{n-1}, a_{n-1}) + R(b_n, a_n)$$

DM: Methods

- Statistical Approaches
  - MDP
    - [Levin et al. 1998, Singh et al. 1999, Levin et al. 2000]
  - POMDP
    - [Roy et al. 2000, Williams and Young 2007]
  - Hidden Information State (HIS)
    - [Young et al. 2010]
  - Bayesian Update of Dialogue State (BUDS)
    - [Thomson 2009]

DM: Recent Trends

- Deep Learning for Dialogue Management
  - Deep Neural Networks for Belief/State Tracking
    - [Henderson et al. 2013, Henderson et al. 2014]

S Hello, How may I help you?

U I need a **Persian** restaurant in the south part of town.

S What kind of food would you like?

U **Persian**.

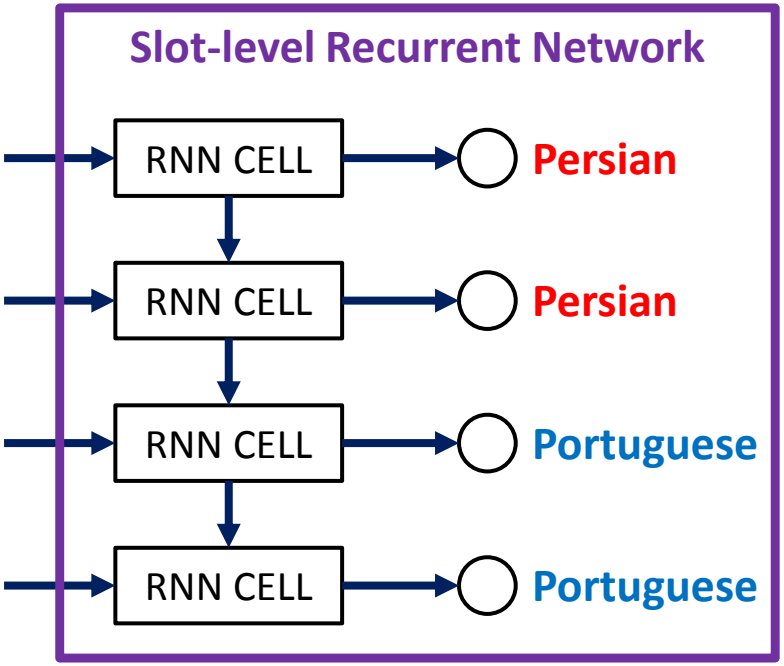
S I'm sorry but there is no restaurant serving persian food

U How about **Portuguese** food?

S Are you looking for Portuguese food?

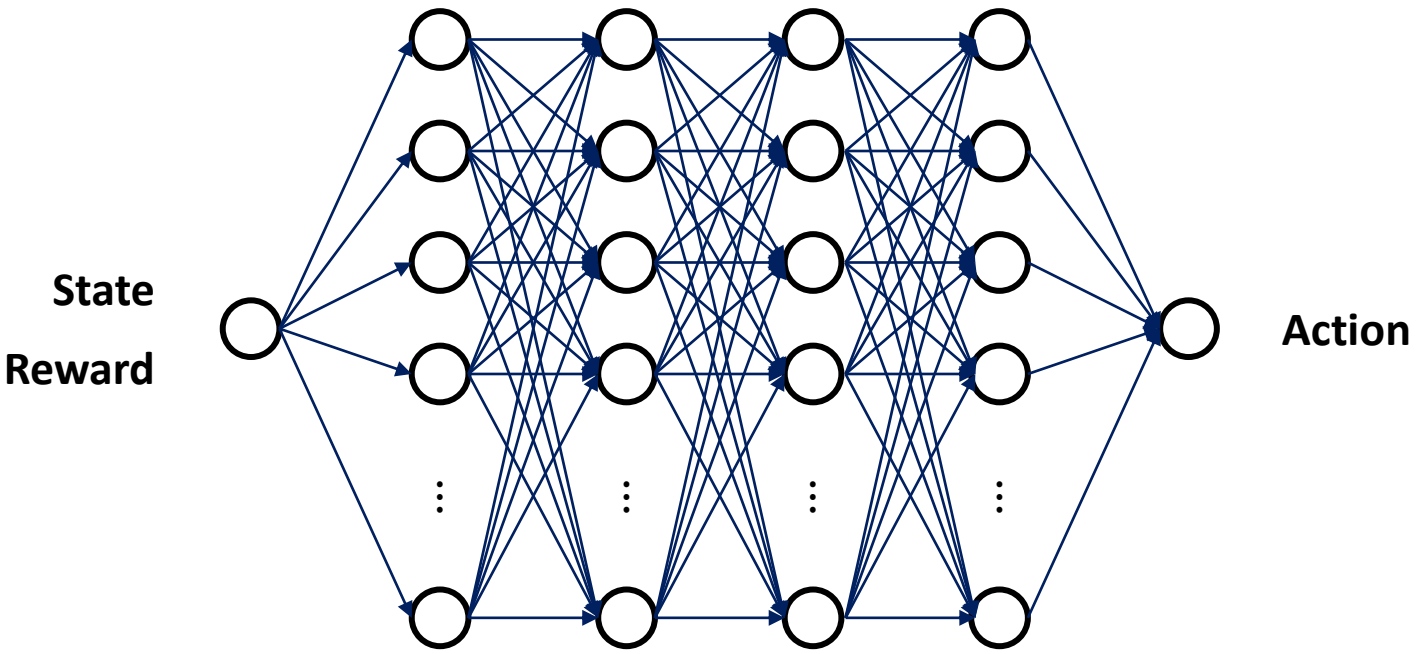
U Yes.

S Nandos is a nice place serving tasty Portuguese food.



DM: Recent Trends

- Deep Learning for Dialogue Management
  - Deep Reinforcement Learning
    - [Cuayáhuitl et al. 2015, Cuayáhuitl 2016]



DM: Recent Trends

- Dialog State Tracking Challenge (DSTC)
  - [Williams et al. 2013, Henderson et al. 2014a, Henderson et al. 2014b, Kim et al. 2016a, Kim et al. 2016b]

Challenge	Type	Domain	Data Provider	Main Theme
DSTC1	Human-Machine	Bus route	CMU	Evaluation metrics
DSTC2	Human-Machine	Restaurant	U. Cambridge	User goal changes
DSTC3	Human-Machine	Tourist information	U. Cambridge	Domain adaptation
DSTC4	Human-Human	Tourist information	I2R	Human conversation
DSTC5	Human-Human	Tourist information	I2R	Language adaptation
DSTC6	In preparation			

  - Mailing list
    - Send an email to listserv@lists.research.microsoft.com
    - With '*subscribe DSTC*' in the body of the message (without quotes)

DM: References

- M. McTear, Modelling Spoken Dialogues With State Transition Diagrams: Experiences With The CSLU Toolkit. in Proceedings of the Fifth International Conference on Spoken Language Processing, 1998.
- R. Pieraccini and J. Huerta. "Where do we go from here? Research and commercial spoken dialog systems." in Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue. 2005.
- D. Traum and S. Larsson. "The information state approach to dialogue management." Current and new directions in discourse and dialogue. Springer Netherlands, 2003. 325-353.

- C. Lee, S. Jung, S. Kim, G. G. Lee. Example-based dialog modeling for practical multi-domain dialog system. Speech Communication,51(5), 466-484, 2009.

- E. Levin, R. Pieraccini, and W. Eckert. "Using Markov decision process for learning dialogue strategies." in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1998.
- S. Singh, M. Kearns, D. Litman, and M. Walker, Reinforcement Learning for Spoken Dialogue Systems. In Proceedings of NIPS 1999.
- E. Levin, R. Pieraccini, and W. Eckert. "A stochastic model of human-machine interaction for learning dialog strategies." IEEE Transactions on speech and audio processing 8.1 (2000): 11-23.

- N. Roy, J. Pineau, and S. Thrun. "Spoken dialogue management using probabilistic reasoning." In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000.
- J. Williams and S. Young. "Partially observable Markov decision processes for spoken dialog systems." Computer Speech & Language 21.2 (2007): 393-422.
- B. Thomson. Statistical methods for spoken dialogue management. Ph.D. thesis, University of Cambridge, 2009.

- S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu, The hidden information state model: A practical framework for POMDP-based spoken dialogue management. Computer Speech & Language, 24(2), 150-174. 2010.

DM: References

- M. Henderson, B. Thomson, and S. Young. "Deep neural network approach for the dialog state tracking challenge." Proceedings of the SIGDIAL 2013 Conference. 2013.
- M. Henderson, B. Thomson, and S. Young. "Word-based dialog state tracking with recurrent neural networks." Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL). 2014.

- H. Cuayáhuitl, S. Keizer, O. Lemon. Strategic Dialogue Management via Deep Reinforcement Learning. in Proceedings of the NIPS Workshop on Deep Reinforcement Learning, 2015.
- H. Cuayáhuitl. SimpleDS: A Simple Deep Reinforcement Learning Dialogue System. in Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS), 2016.

- J. Williams, A. Raux, D. Ramachandran, and A. Black, “The dialog state tracking challenge,” in Proceedings of the SIGDIAL 2013 Conference, 2013, pp.404–413.
- M. Henderson, B. Thomson, and J. Williams, “The second dialog state tracking challenge,” in 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2014, p. 263.
- M. Henderson, B. Thomson, and J. Williams, “The third dialog state tracking challenge,” in Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014, pp. 324–329.
- S. Kim, L. F. D’Haro, R. E. Banchs, J. Williams, and M. Henderson, “The Fourth Dialog State Tracking Challenge,” in Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS), 2016.
- S. Kim, L. F. D’Haro, R. E. Banchs, J. Williams, M. Henderson, and K. Yoshino, “The Fifth Dialog State Tracking Challenge,” in Proceedings of the 2016 IEEE Workshop on Spoken Language Technology (SLT), 2016.

DM: Resources

- Datasets
  - Communicator
    - https://catalog.ldc.upenn.edu/LDC2004T16
  - CMU
    - http://www.speech.cs.cmu.edu/letsgo/letsgodata.html
  - University of Cambridge
    - http://mi.eng.cam.ac.uk/research/dialogue/corpora/
    - http://camdial.org/~mh521/dstc/
  - Ubuntu Dialogue Corpus
    - https://github.com/rkadlec/ubuntu-ranking-dataset-creator
  - TourSG
    - http://www.colips.org/workshop/dstc4/data.html
- Toolkits
  - Ravenclaw
    - http://wiki.speech.cs.cmu.edu/olympus/index.php/RavenClaw
  - TrindiKit
    - http://www.ling.gu.se/projekt/trindi/trindikit/
  - OpenDial
    - http://www.opendial-toolkit.net/

SUMMARY

- Natural language understanding aims to interpret user intention in natural language to domain-specific semantic representations
- Dialogue manager monitors belief states at each turn in dialogues and determines the system action accordingly
- Statistical approaches have been preferred to build both the components compared to conventional knowledge or rule-based approaches
- Recently, deep neural network models have achieved performance improvements in both tasks

PART 3: SYSTEM COMPONENTS AND ARCHITECTURES

Rafael E. Banchs, Seokhwan Kim, **Luis Fernando D'Haro**, Andreea I. Niculescu

Human Language Technology (I^2R, A*STAR)

HAI 2016
4th International Conference on Human-Agent Interaction
4 - 7 OCTOBER
SINGAPORE

A*STAR
Institute for Infocomm Research

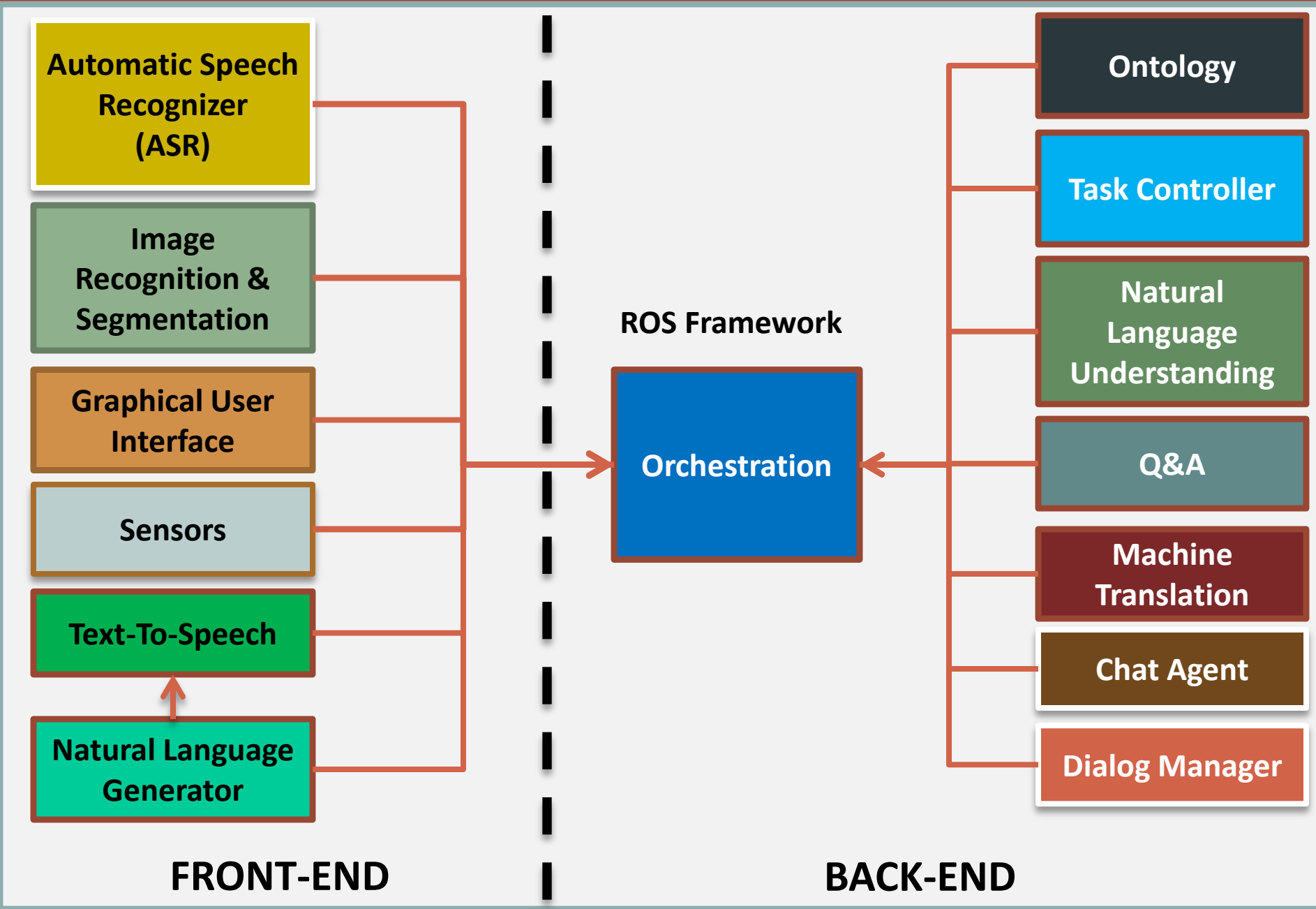
Tutorial Content Overview

1. Natural Language in Human-Robot Interaction
   - *Human-Robot Interaction*
   - *The Role of Natural Language*
2. Semantics and Pragmatics
   - *Natural Language Understanding*
   - *Dialogue Management*
3. System Components and Architectures
   - *Front-end System Components (Interfaces)*
   - *Back-end System Components*
4. User Experience (UX) Design and Evaluation
   - *UX design for Speech Interactions*
   - *User Studies and Evaluations*

Example of Conversational Robots

- **AIBO**: Series of robotic pets designed and manufactured by Sony since 1998 until 2005
  - Built-in Speech Recognition
  - Heat, acceleration, vibration, velocity sensors
  - 20 degrees of freedom
  - Image sensor 350K pixels
- **NAO**: Humanoid robot developed by Aldebaran Robotics (France) since 2004
  - 2 HD cameras,
  - 4 microphones,
  - sonar range finder,
  - 2 infrared emitters and receivers,
  - 9 tactile sensors,
  - 8 pressure sensors
  - Ethernet/Wi-Fi
  - SDK for different languages
- **PEPPER**: Humanoid robot by Aldebaran Robotics and SoftBank (2014)
  - Designed with the ability to read emotions
  - facilitate relationships,
  - have fun with people, and
  - connect people with the outside world

Components of Human-Robot Interfaces

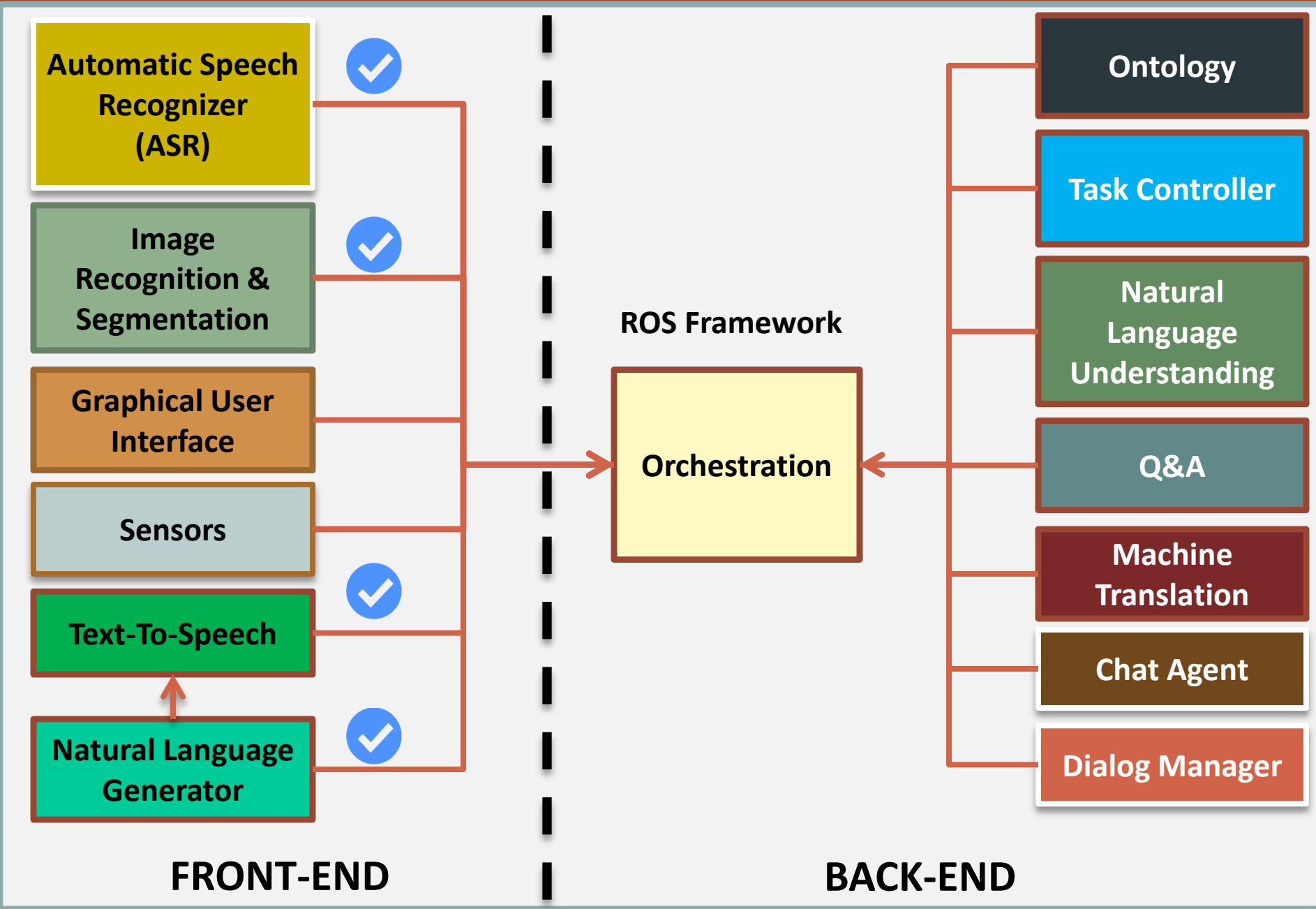


Part 3:
System Components and Architecture

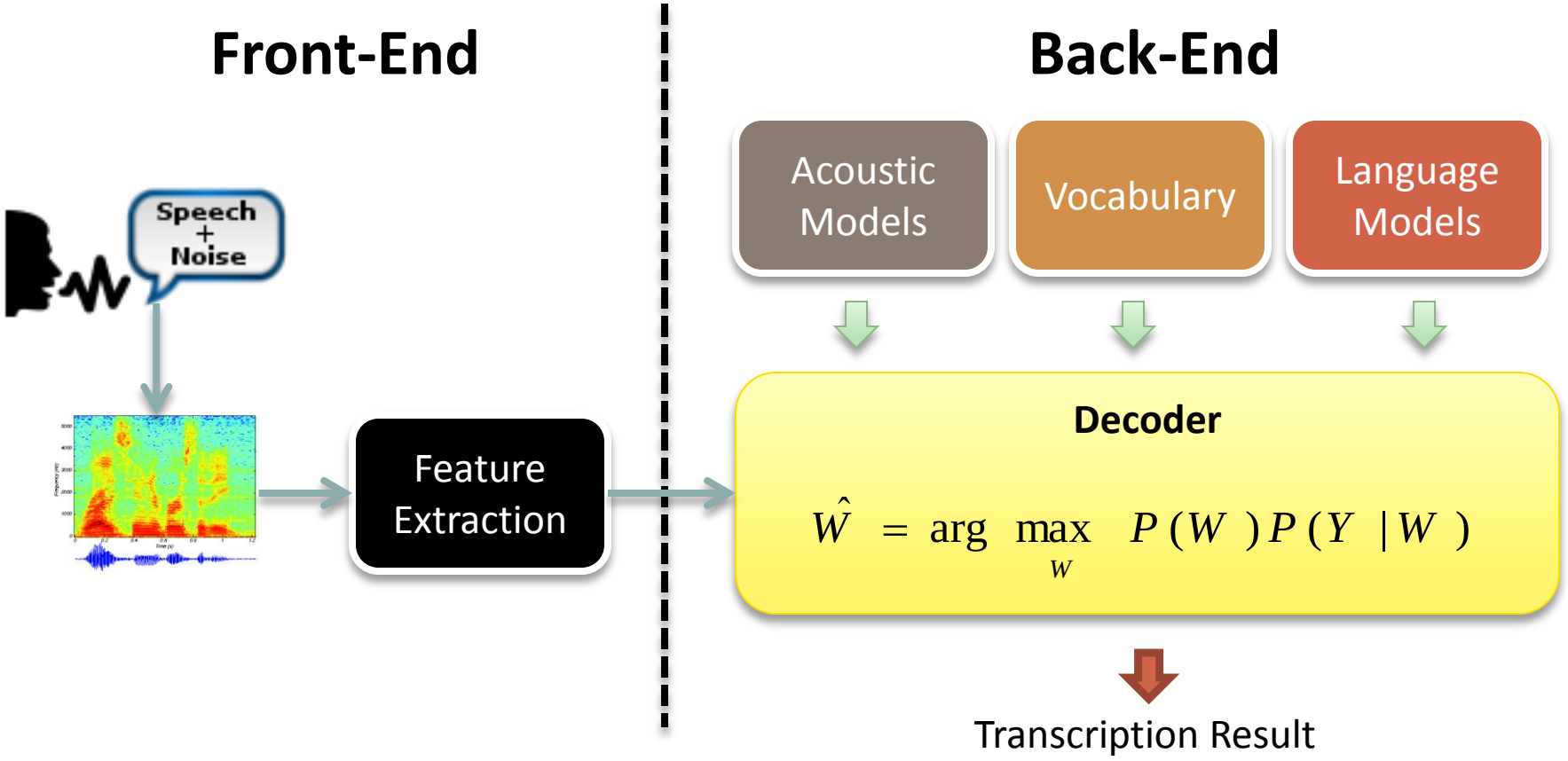
FRONT-END SYSTEM COMPONENTS

Institute for Infocomm Research
A*STAR

Components of Human-Robot Interfaces



Automatic Speech Recognition

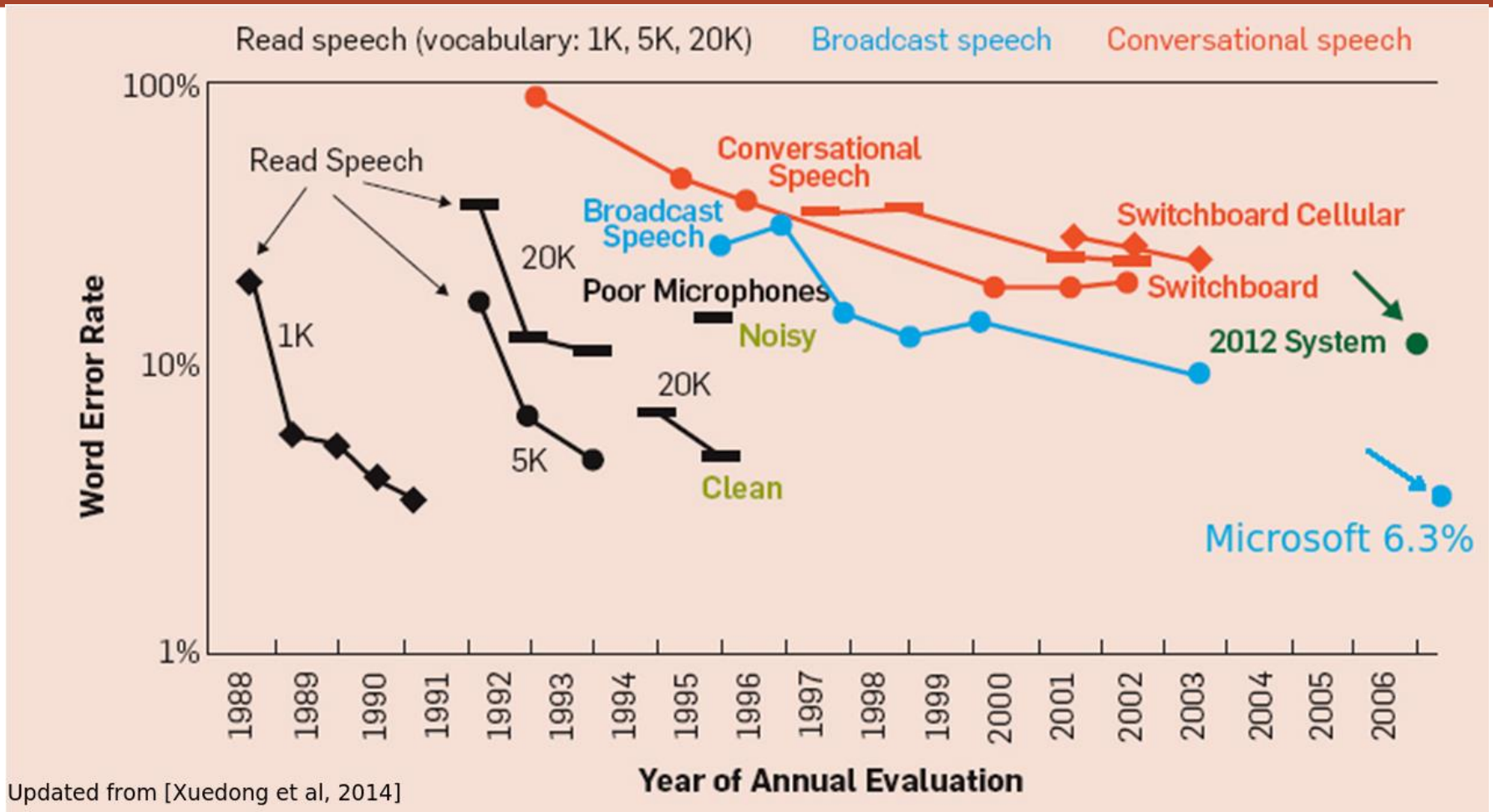


- Front-end: Captures/Pre-process the speech signal
- Back-end: Combines different information and search for optimal sequence

ACOUSTIC & LANGUAGE MODELS + VOCABULARY

- AMs:
  - Models acoustic variabilities & maps sounds to phonemes/words
    - Typical models: HMMs [Lawrence and Rabiner, 1989] and DNN/HMM [Hinton et al, 2012]
- LMs:
  - Models the grammar of the recognized sentence
    - Finite state grammars [Mohri et al, 2002]
    - Statistical models [Chen and Goodman, 1999], [Bengio et al, 2003], [Jozefowicz et al, 2016]
- Vocabulary:
  - Maps probabilities of phone sequence into orthographic representation
    - Conversion rules [Rao et al, 2015]
- Decoder:
  - Maximizes the combination of the different sources of information and provides the final transcription

PROGRESS ON ASR

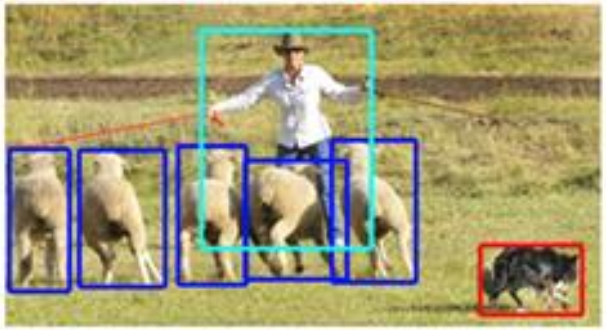


- AMs: Usage of very deep DNN and huge number of acoustic annotated data
- LMs: Word-vector embeddings to capture semantic relationships + RNNs to capture context
- Vocabulary: Handling multiple variations + dialects

Image Recognition and Segmentation

- Allows detection and recognition of objects, people, emotions, face tracking
- Deep Complex architectures based mainly on using Convolutional Neural Networks
- Several number of well-trained models are available
  - Fine-tune is required for particular objects or tasks
- Combined with syntactic parsing could be used for object location and path planning [Gutierrez et al, 2015]

(a) classification

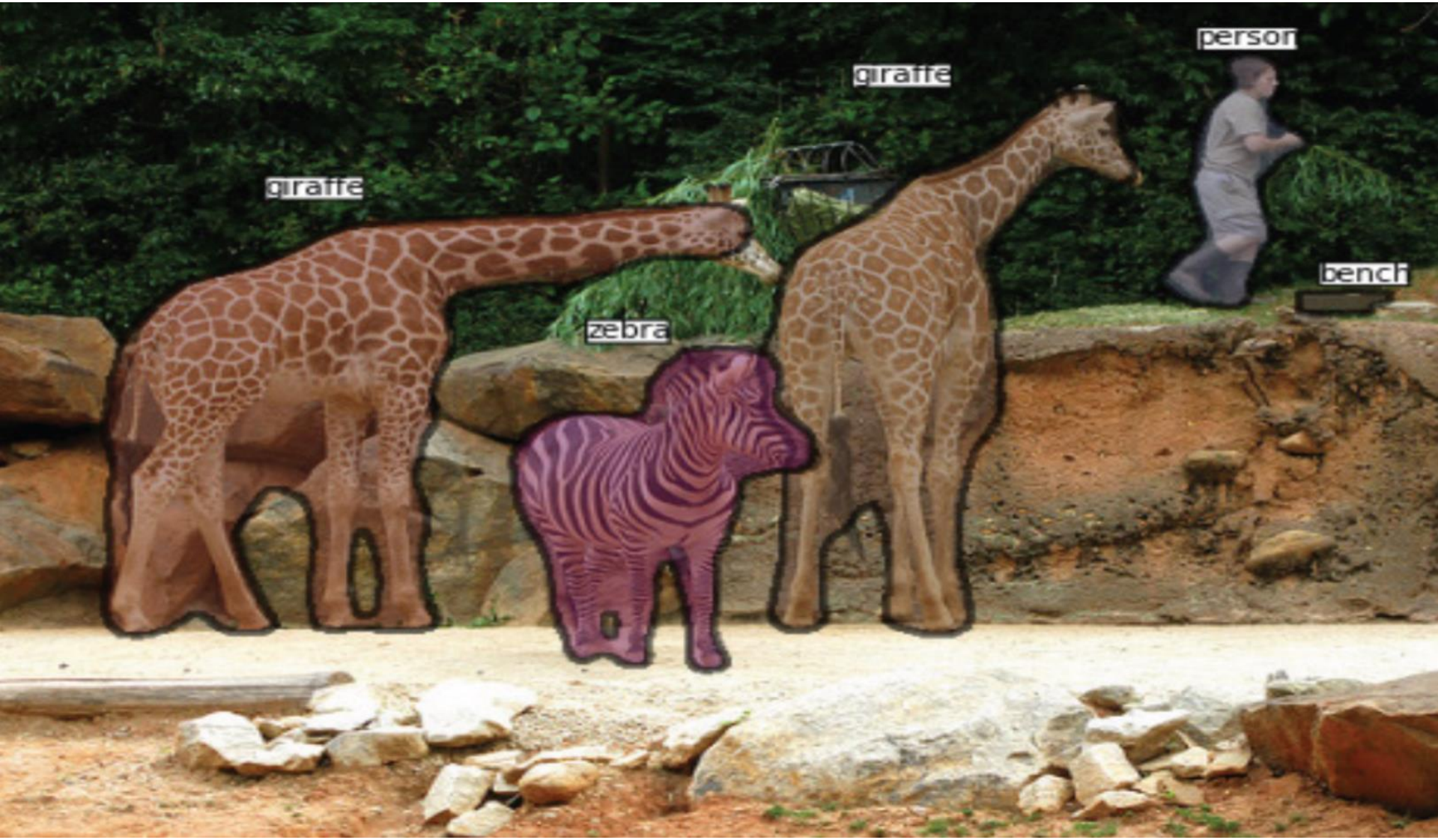


(b) detection

(c) segmentation

Image from https://code.facebook.com/posts/561187904071636/segmenting-and-refining-images-with-sharpmask/

Example of Image Recognition and Segmentation



Zagoruyko, S., Lerer, A., Lin, T. Y., Pinheiro, P. O., Gross, S., Chintala, S., & Dollár, P. (2016). A MultiPath Network for Object Detection. *arXiv preprint arXiv:1604.02135*.

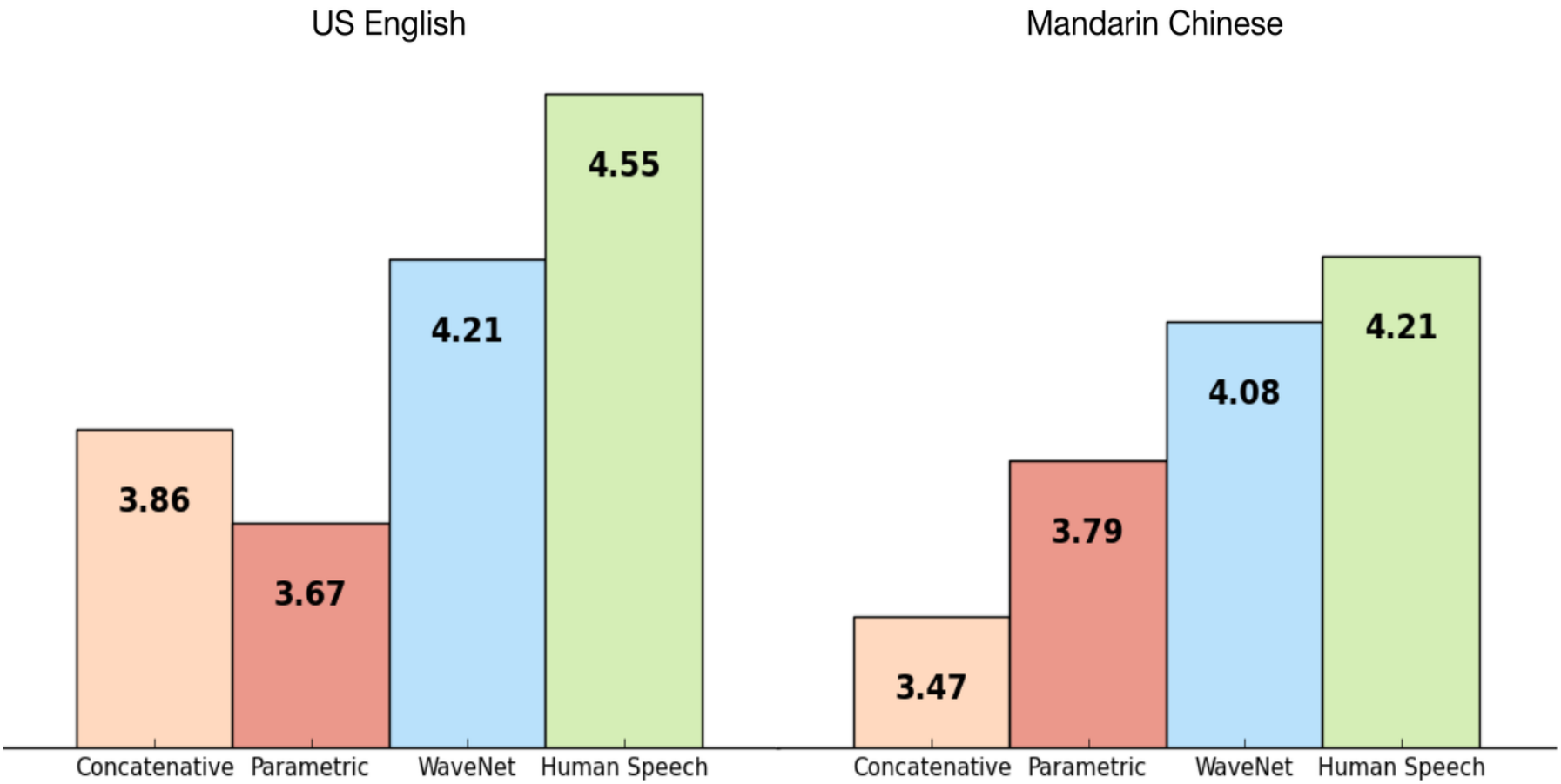
Available at https://github.com/facebookresearch/multipathnet

Text-To-Speech Conversion

- Useful for providing attention messages to the users
  - Specially relevant for interaction with kids and blind people
- Main approaches
  - Unit-selection [Hunt et al, 1996]: Requires long number of recordings at different levels (phrases, words, phonemes, tri-phonemes, etc.) that are concatenated to produce the sound
  - Parametric [Klatt, 1987]: Modification of parameters send to a vocoder which produces the sounds
  - Generative model [van den Oord, 2016]: Learn to generate each audio sample, e.g. WaveNet

Subjective Evaluation TTS

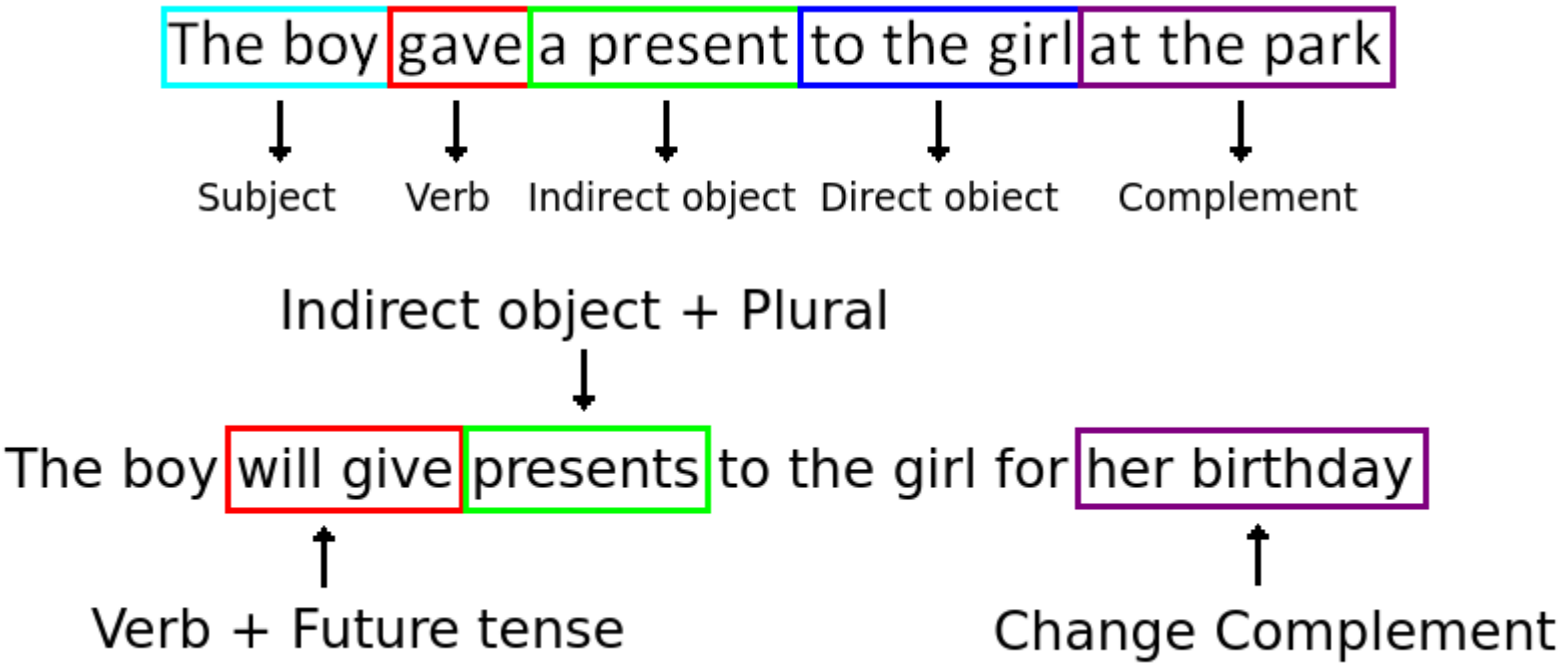
- Parametric Concatenative Generative



- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. WAVENET: a generative model for raw audio. https://arxiv.org/pdf/1609.03499.pdf

Other Components

- NLG: Natural Language Generation
  - Allows creating new sentences by means of configurable templates
  - Parameters allows adaptation to context (e.g. tense, gender, number, style, etc.)
  - E.g.



Other Components

- Sound localization using arrays of microphones
  - Detect which person is speaking (a.k.a. speaker id + diarization)
  - Improves accuracy of ASR
  - Allows showing attention (specially when combined with face tracking)

Main References

- Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. journal of machine learning research, 3(Feb), 1137-1155.
- Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. Computer Speech & Language, 13(4), 359-394.
- Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 30-42.
- D'Haro, L. F., & Banchs, R. E. (2016). Automatic Correction of ASR outputs by Using Machine Translation, in Proceedings Interspeech 2016, pps. 3469-3473
- Gutierrez, M. A., Banchs, R. E., & D'Haro, L. F. Perceptive Parallel Processes Coordinating Geometry and Texture, in Proceedings of the Workshop on Multimodal and Semantics for Robotics Systems (MuSRobS) co-located with IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2015), pp 30-35, Hamburg, Germany, October 1, 2015
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.
- Hunt, A. J., & Black, A. W. (1996, May). Unit selection in a concatenative speech synthesis system using a large speech database. In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on (Vol. 1, pp. 373-376). IEEE.
- Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410.
- Klatt, D. H. (1987). Review of text-to-speech conversion for English. The Journal of the Acoustical Society of America, 82(3), 737-793.

Main References

- Lawrance, R., & Rabiner, A. (1989). Tutorial on hidden Markov models and selected application in speech recognition. Proceedings of the IEEE, 77(2), 257-286.
- Mohri, M., Pereira, F., & Riley, M. (2002). Weighted finite-state transducers in speech recognition. Computer Speech & Language, 16(1), 69-88.
- Rao, K., Peng, F., Sak, H., & Beaufays, F. (2015, April). Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4225-4229). IEEE.
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kavukcuoglu, K. Wavenet: a Generative model For RAW audio. https://arxiv.org/pdf/1609.03499.pdf
- Xuedong Huang, James Baker, Raj Reddy "A Historical Perspective of Speech Recognition", in Communications of the ACM, January 2014 Vol. 57 No. 1, Pages 94-103, DOI: 10.1145/2500887
- Zagoruyko, S., Lerer, A., Lin, T. Y., Pinheiro, P. O., Gross, S., Chintala, S., & Dollár, P. (2016). A MultiPath Network for Object Detection. arXiv preprint arXiv:1604.02135.

Additional References

- NLG:
  - Paris, C., Swartout, W. R., & Mann, W. C. (Eds.). (2013). Natural language generation in artificial intelligence and computational linguistics (Vol. 119). Springer Science & Business Media.
  - McDonald, D. D. (2010). Natural Language Generation. Handbook of natural language processing, 2, 121-144.
  - Reiter, E., Dale, R., & Feng, Z. (2000). Building natural language generation systems (Vol. 33). Cambridge: Cambridge university press.
- Array of Microphones
  - Pavlidi, D., Griffin, A., Puigt, M., & Mouchtaris, A. (2013). Real-time multiple sound source localization and counting using a circular microphone array. IEEE Transactions on Audio, Speech, and Language Processing, 21(10), 2193-2206.
  - Valin, J. M., Michaud, F., Rouat, J., & Létourneau, D. (2003, October). Robust sound source localization using a microphone array on a mobile robot. In Intelligent Robots and Systems, 2003.(IROS 2003). Proceedings. 2003 IEEE/RSJ International Conference on (Vol. 2, pp. 1228-1233). IEEE.
- Diarization
  - Liu, Y., Tian, Y., He, L., & Liu, J. (2016). Investigating Various Diarization Algorithms for Speaker in the Wild (SITW) Speaker Recognition Challenge. Interspeech 2016}, 853-857.
  - Tranter, S. E., & Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech, and Language Processing, 14(5), 1557-1565.

RESOURCES

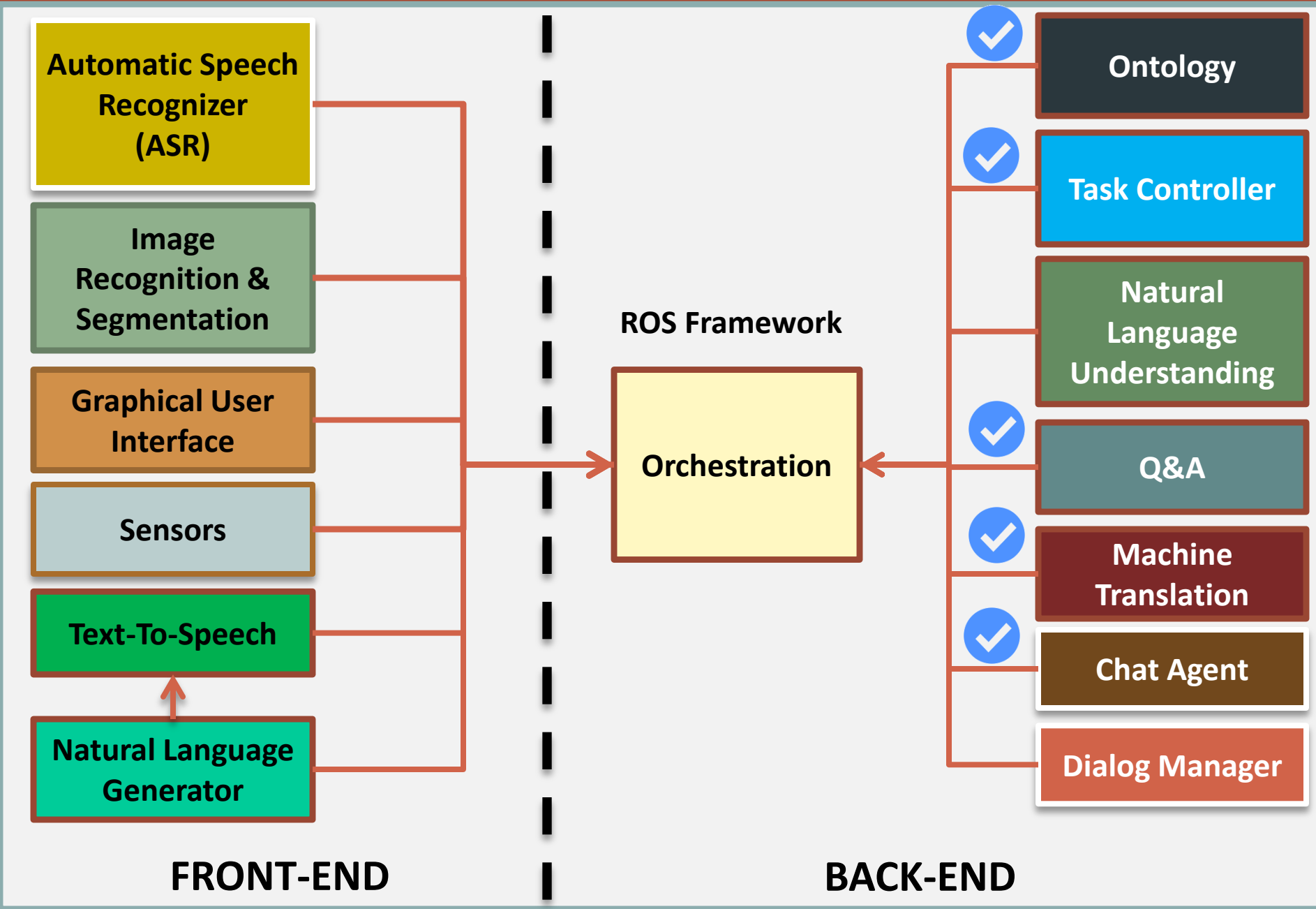
- ASR toolkits
  - Kaldi: http://kaldi-asr.org/
  - Sphinx: http://cmusphinx.sourceforge.net/
  - HTK: http://htk.eng.cam.ac.uk/links/asr_tool.shtml
  - RWTH: https://www-i6.informatik.rwth-aachen.de/rwth-asr/
- TTS toolkits
  - HTS: http://hts.sp.nitech.ac.jp/
  - Wavenet: https://deepmind.com/blog/wavenet-generative-model-raw-audio/
- Image Recognition and Description
  - Multipath: https://github.com/facebookresearch/multipathnet
  - Inception: https://github.com/tensorflow/models/tree/master/inception
- NLG
  - RNNLG: https://github.com/shawnwun/RNNLG
  - SimpleNLG: https://github.com/simplenlg/simplenlg
- Speaker recognition and diarization:
  - Spear: https://pythonhosted.org/bob.bio.spear/
  - Alize: http://mistral.univ-avignon.fr/
  - Sidekit: https://pypi.python.org/pypi/SIDEKIT

Part 3:

System components and Architecture

BACK-END SYSTEM COMPONENTS

Components of Human-Robot Interfaces



THE BOT PLATFORM ECOSYSTEM

- Several AI agents
  - Siri, Google Now, Cortana, Alexa/Echo
- Messaging platforms:
  - Facebook Messenger, Telegram, WebChat, Slack, Skype
- AI service platforms:
  - Google cloud, IBM Watson, LUIS, Deepmind
- Bots:
  - Api.ai, pandorabots, Automat, Bot Framework

Image from Jon Bruner - https://www.oreilly.com/ideas/infographic-the-bot-platform-ecosystem

Conversational Agents (Bots/Chatbots)

- Used mainly to allow non-directed conversations
  - Jokes: [Devillers and Soury, 2013]
  - Chat: [Jokinen and Wilcock, 2014]
- Main approaches:
  - Rule-based systems, e.g.
    - Eliza [Weizenbaum, 1966], ALICE
  - Retrieval-based: predefined responses + heuristics to pick response based on the input and context, e.g.
    - IRIS [Banchs et al, 2012], Cleverbot
  - Generative models: system generates the answer from scratch
    - Seq2Seq [Vinyals et al, 2015]

CHATTING ABOUT GENERAL KNOWLEDGE

Human: *What do you think about messi ?*
Machine: *he 's a great player .*
Human: *what do you think about cleopatra ?*
Machine: *oh , she 's very regal .*
Human: *what do you think about england during the reign of elizabeth ?*
Machine: *it was a great place .*

- Notice the correct use of pronouns

Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).

TAY: A MISLEAD CONVERSATIONAL AGENT

- Released by Microsoft on Mar 23, 2016
- Learned to provide answers by adapting its KB from the interactions with people on Tweeter

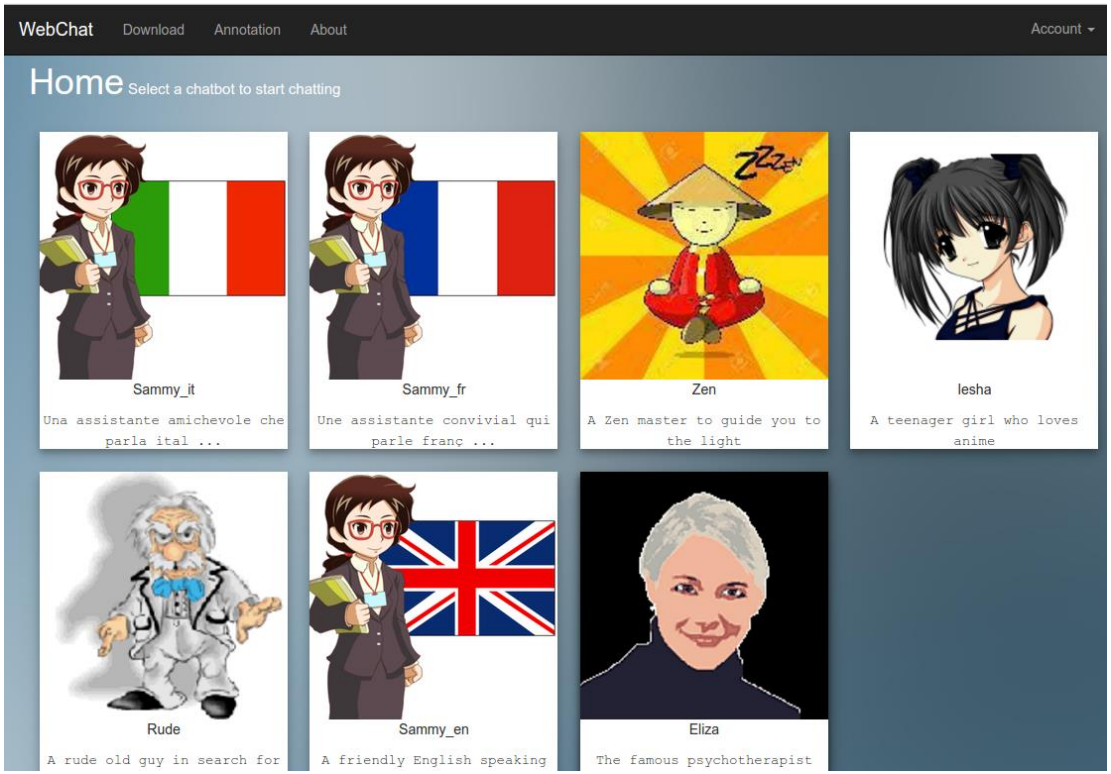
Institute for Infocomm Research

Available Datasets

- Annotated dialogs are required to train reliable chatbots
- Different levels of annotations:
  - Adequacy, correctness, politeness, context-aware
- Available datasets: [Serban et al, 2015]
  - Ubuntu dialog, [Lowe et al, 2015]
  - MovieDic [Banchs & Li, 2012],
  - TourSG [Kim et al, 2016]
  - Reddit [Al-Rfou et al, 2016]

Online Data Collection

- WebChat proposed by [Lin et al, 2016]
- API interface to connect any kind of chatbot
- Different user profiles
  - Annotators, chatbot providers, and users
- Annotation forms
- Gammification
  - Downloading based on earned points



13 chatbots, more than 11K turns collected, around 3K annotated sentences!!

Lue Lin, Luis F. D'Haro, Rafael Banchs. A Web-based Platform for Collection of Human-Chatbot Interactions. To appear in Proceedings IV International Conference on Human-Agent Interaction, HAI'16. Singapore, October 4-7, 2016.
http://www.teachabot.com:8000/main

Main Challenges & Directions

- Consistency: Keep similar answers in spite of different wordings[1]

Human: *what is your job ?*
Machine: *i 'm a lawyer .*
Human: *what do you do ?*
Machine: *i 'm a doctor .*

- Quick domain-dependent adaptation: Specially from un-structured data[2]
- Personalization: Handling profiles, interaction levels, and keep relevant context history
- Long sentence generation: most sentence are short or common phrases

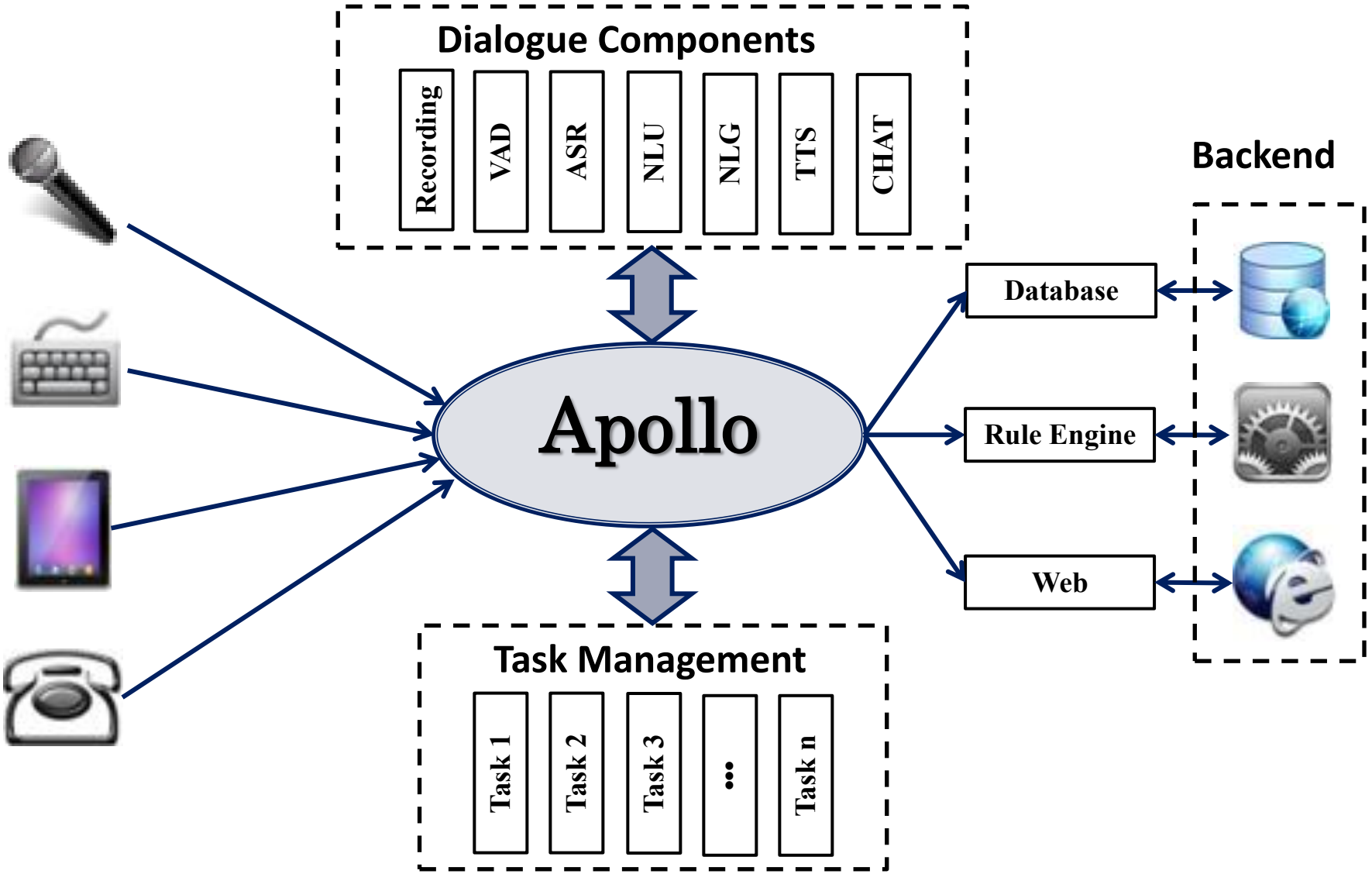
[1] Example take n from Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).
[2] Yan, Z., Duan, N., Bao, J., Chen, P., Zhou, M., Li, Z., & Zhou, J. DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents. ACL 2016, Berlin, Germany.

FRAMEWORKS

- Allows integration and orchestration of the different components required to perform the dialog, especially relevant:
  - Definition of NLU grammars, ASR vocabulary and models, execution of tasks, flow logic
- Example of well known frameworks
  - Olympus from CMU [Bohus & Rudnicky, 2009]
  - OpenDial from University of Oslo [Lison & Kennington, 2016]
  - Trindikit from University of Gothenburg [Larsson and Traum, 2000]
  - Apollo from A*STAR [Jiang et al, 2014]
    - Used in [Gutierrez et al, 2016] -- Paper at HAI'16
    - SERC industrial project (EC-2013-045)[1]

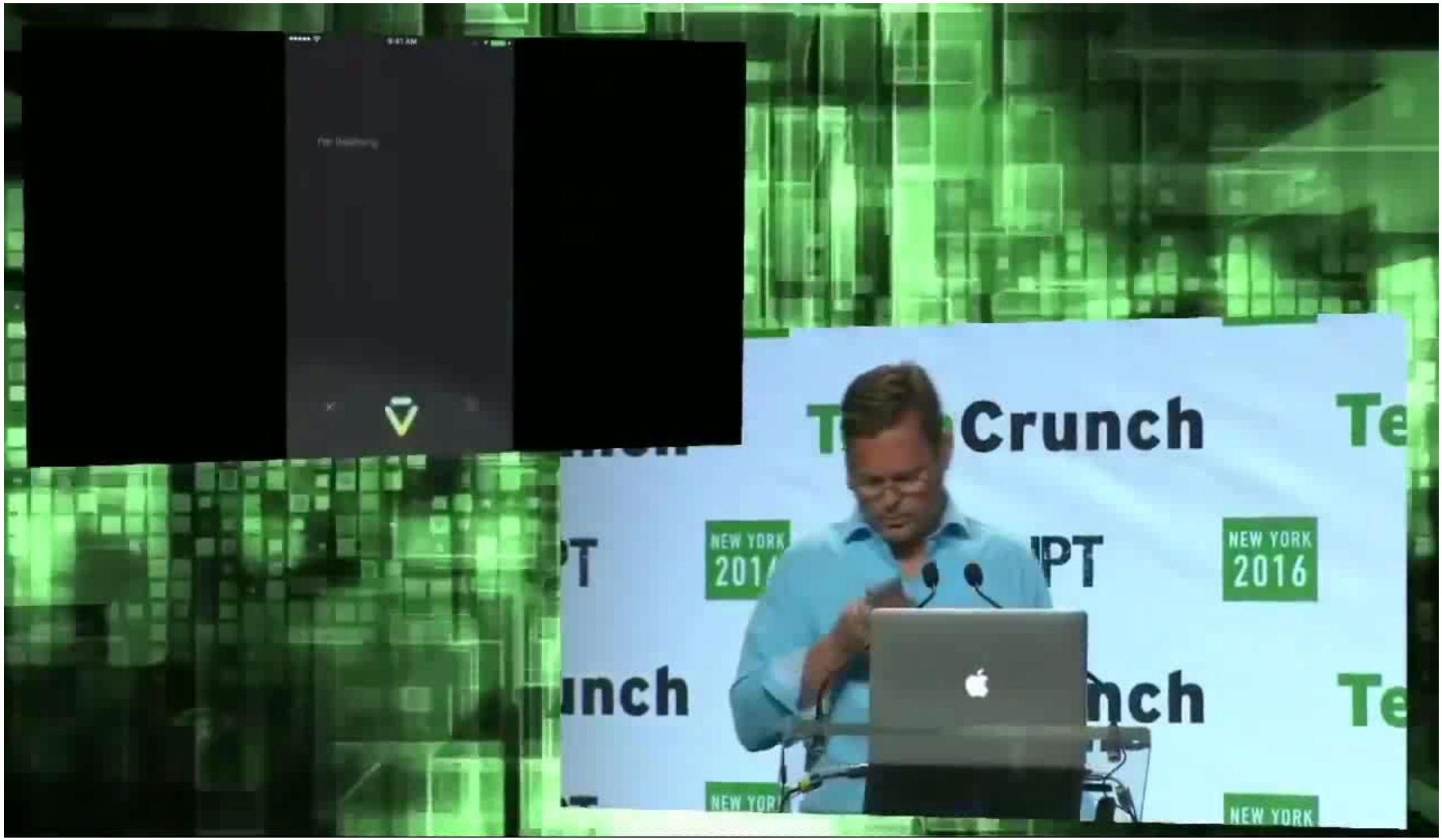
[1] http://www.singaporerobotics.org

Apollo Framework



Jiang, R., Tan, Y. K., Limbu, D. K., Dung, T. A., & Li, H. (2014). Component pluggable dialogue framework and its application to social robots. In Natural Interaction with Robots, Knowbots and Smartphones (pp. 225-237). Springer New York.

Understanding Complex Questions

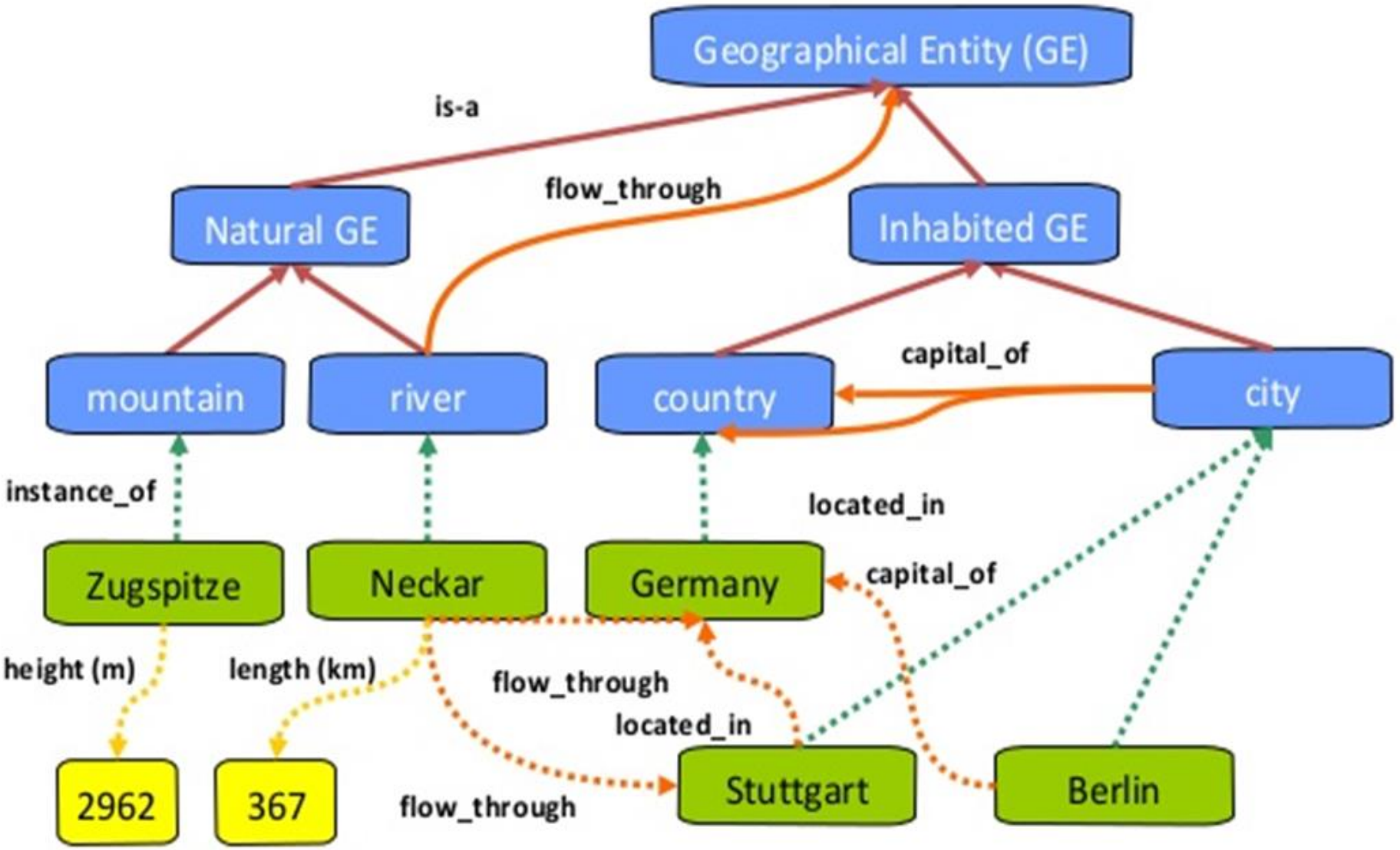


Presentation of IVY by Dag Kittlaus, May2016. Available at https://www.youtube.com/watch?v=Rblb3sptgpQ

Other Components

- Task Controller
  - Transform orders into actual low-level commands for the robot
- Ontologies
  - Allows knowledge from the domain and world
  - Keeps information about relations between entities, their types and properties
- Machine Translation:
  - Used to handle multi-lingual capabilities without changing the logic inside
  - For an industrial robot: Used to correct errors from the ASR or adapt to a domain [D'Haro et al, 2016]

Ontology Example



Taken from http://www.slideshare.net/SanthoshKannan/4-semantic-web-and-ontology

Other Components

- Q&A:
  - Complementary to the Ontology, allows answering questions about the robot, its operation, capabilities, as well as world/task knowledge
  - Typically based on using indexes (Lucene), databases (MySQL), or knowledge graphs (SparQL)

VQA with Attention Mechanism

What is the color of the coat?

Traditional VQA: analyze the whole image -> analyze question -> give answer: ~~brown~~

Attention based VQA: find coat -> judge the color of coat -> give answer: **yellow**

What is the color of the umbrella?

Traditional VQA: analyze the whole image -> analyze question -> give answer: ~~green~~

Attention based VQA: find umbrella -> judge the color of umbrella -> give answer: **red**

Chen, K., Wang, J., Chen, L. C., Gao, H., Xu, W., & Nevatia, R. (2015). ABC-CNN: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960.

Main References

- Al-Rfou, R., Pickett, M., Snaider, J., Sung, Y. H., Strope, B., & Kurzweil, R. (2016). Conversational Contextual Cues: The Case of Personalization and History for Response Ranking. arXiv preprint arXiv:1606.00372.
- Banchs, R. E., & Li, H. (2012, July). IRIS: a chat-oriented dialogue system based on the vector space model. In Proceedings of the ACL 2012 System Demonstrations (pp. 37-42). Association for Computational Linguistics.
- Banchs, R. E. (2012, July). Movie-DiC: a movie dialogue corpus for research and development. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2 (pp. 203-207). Association for Computational Linguistics.
- Bohus, D., & Rudnicky, A. I. (2009). The RavenClaw dialog management framework: Architecture and systems. Computer Speech & Language, 23(3), 332-361.
- Chen, K., Wang, J., Chen, L. C., Gao, H., Xu, W., & Nevatia, R. (2015). ABC-CNN: An attention based convolutional neural network for visual question answering. arXiv preprint arXiv:1511.05960.
- D'Haro, L. F. and Lin, L. An Online Platform for crowd-Sourcing Data from Interactions with Chatbots. WOCHAT shared-task report, Intelligent Virtual Agents (IVA 2016), September 20-23, 2016, Los Angeles, California
- Devillers, L. Y., & Soury, M. (2013, December). A social interaction system for studying humor with the robot nao. In Proceedings of the 15th ACM on International conference on multimodal interaction (pp. 313-314). ACM.
- Gutierrez, M.A., D'Haro, L. F., and Banchs, R. E. (2016) A Multimodal Control Architecture for Autonomous Unmanned Aerial Vehicles. To appear in proceedings IV International Conference on Human Agent-Interaction, Singapore, Oct 4-7, 2016.
- Jiang, R., Tan, Y. K., Limbu, D. K., Dung, T. A., & Li, H. (2014). Component pluggable dialogue framework and its application to social robots. In Natural Interaction with Robots, Knowbots and Smartphones (pp. 225-237). Springer New York.

Main References

- Jokinen, K., & Wilcock, G. (2014). Multimodal open-domain conversations with the Nao robot. In Natural Interaction with Robots, Knowbots and Smartphones (pp. 213-224). Springer New York.
- Kim, S., D'Haro, L. F., Banchs, R. E., Williams, J. D., & Henderson, M. (2016). The fourth dialog state tracking challenge. In Proceedings of the 7th International Workshop on Spoken Dialogue Systems (IWSDS). Information about corpus at http://workshop.colips.org/dstc5/data.html
- Larsson, Staffan and Traum, David (2000): Information state and dialogue management in the TRINDI Dialogue Move Engine Toolkit. In Natural Language Engineering Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering, Cambridge University Press, U.K. (pp. 323-340, 18 pages)
- Lue Lin, Luis F. D'Haro, Rafael Banchs. A Web-based Platform for Collection of Human-Chatbot Interactions. To appear in Proceedings IV International Conference on Human-Agent Interaction, HAI'16. Singapore, October 4-7, 2016.
- Lison, P., & Kennington, C. (2016). OpenDial: A Toolkit for Developing Spoken Dialogue Systems with Probabilistic Rules. ACL 2016, 67.
- Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909.
- Serban, I. V., Lowe, R., Charlin, L., & Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. arXiv preprint arXiv:1512.05742.
- Vinyals, Oriol, and Quoc Le. "A neural conversational model." arXiv preprint arXiv:1506.05869 (2015).
- Yan, Z., Duan, N., Bao, J., Chen, P., Zhou, M., Li, Z., & Zhou, J. DocChat: An Information Retrieval Approach for Chatbot Engines Using Unstructured Documents. ACL 2016, Berlin, Germany.
- Weizenbaum, Joseph (1966). "ELIZA—a computer program for the study of natural language communication between man and machine". Communications of the ACM. New York, NY: Association for Computing Machinery. 9 (1): 36–45.

Additional References

- Corpus
  - Ameixa, D., Coheur, L., & Redol, R. A. (2013). From subtitles to human interactions: introducing the subtle corpus. Tech. rep., INESC-ID (November 2014).
  - Banchs, R. E. 2012. Movie-DiC: A movie dialogue corpus for research and development. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 203–207.
  - Levesque, H. J., Davis, E., & Morgenstern, L. (2011, March). The Winograd schema challenge. In AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning (Vol. 46, p. 47).
  - Lowe, R., Pow, N., Serban, I., & Pineau, J. (2015). The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. arXiv preprint arXiv:1506.08909.
  - Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., & Vanderwende, L. (2016). Generating Natural Questions About an Image. arXiv preprint arXiv:1603.06059.
  - Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv preprint arXiv:1606.05250.
  - Serban, I. V., Lowe, R., Charlin, L., & Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. arXiv preprint arXiv:1512.05742.

- Reasoning
  - Kumar, G., Banchs, R. E., & D'Haro, L. F. (2015, October). Automatic fill-the-blank question generator for student self-assessment. In Frontiers in Education Conference (FIE), 2015. 32614 2015. IEEE (pp. 1-3). IEEE.
  - Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., & Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698.

Additional References

- Q&A:
  - Allam, A.M. N., & Haggag, M. H. (2012). The question answering systems: A survey. International Journal of Research and Reviews in Information Sciences (IJRRIS), 2(3).
  - Gupta, P., & Gupta, V. (2012). A survey of text question answering techniques. International Journal of Computer Applications, 53(4).
  - Iyyer, M., Boyd-Graber, J. L., Claudino, L. M. B., Socher, R., & Daumé III, H. (2014). A Neural Network for Factoid Question Answering over Paragraphs. In EMNLP (pp. 633-644).
- Ontologies
  - Staab, S., & Studer, R. (Eds.). (2013). Handbook on ontologies. Springer Science & Business Media.
  - Van Harmelen, F., Lifschitz, V., & Porter, B. (Eds.). (2008). Handbook of knowledge representation (Vol. 1). Elsevier.
- Machine Translation
  - Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
  - Koehn, P. (2009). Statistical machine translation. Cambridge University Press.

RESOURCES

- Corpus:
  - UBUNTU: http://cs.mcgill.ca/~jpineau/datasets/ubuntu-corpus-1.0/
  - TRAINS: https://www.cs.rochester.edu/research/speech/trains.html
  - MovieDIC: [Banchs, 2012]
  - Subtle: [Ameixa et al, 2014]
  - SQuAD: [Rajpurkar et al, 2016]
  - OpenSubtitles: http://www.opensubtitles.org/

- Chatbot frameworks
  - https://chatfuel.com/ (Not coding at all)
  - https://howdy.ai/botkit/ (chatbot toolkit for slack)
  - https://dev.botframework.com/ (from Microsoft)

- Reasoning datasets
  - BABI [Weston et al, 2015]
  - Winograd scheme [Levesque et al, 2011]

RESOURCES

- Ontologies:
  - Protégé: http://protege.stanford.edu/
- Q&A/VQ&A:
  - QANTA: https://cs.umd.edu/~miyyer/qblearn/
  - Show and tell: https://github.com/tensorflow/models/tree/master/im2txt
- Machine Translation
  - Moses: http://www.statmt.org/moses/
  - Seq2Seq: https://www.tensorflow.org/versions/r0.10/tutorials/seq2seq/index.html

SUMMARY

- Human-Robot interfaces is a hot topic
  - However, several components must be integrated
  - Existing frameworks can be used and connected using ROS
- Most state-of-the-art technologies are based on Deep Neural Networks
  - Requires huge amounts of labeled data
  - Several frameworks/models are available
- Main challenges:
  - Fast domain adaptation with scarse data + re-use of rules/knowledge
  - Handling reasoning
  - Data collection and analysis from un-structured data
  - Complex-cascade systems requires high accuracy for working good as a whole

PART 4: USER EXPERIENCE DESIGN AND EVALUATION

Rafael E. Banchs, Seokhwan Kim, Luis Fernando D'Haro, **Andreea I. Niculescu**

Human Language Technology (I^2R, A*STAR)

HAI 2016 4 - 7 OCTOBER SINGAPORE

4th International Conference on Human-Agent Interaction

A*STAR Institute for Infocomm Research

Tutorial Content Overview

1. Natural Language in Human-Robot Interaction
   - *Human-Robot Interaction*
   - *The Role of Natural Language*
2. Semantics and Pragmatics
   - *Natural Language Understanding*
   - *Dialogue Management*
3. System Components and Architectures
   - *Front-end System Components (Interfaces)*
   - *Back-end System Components*
4. User Experience Design (UX) and Evaluation
   - *UX Design for Speech Interactions*
   - *User Studies and Evaluations*

Part 4:
User Experience Design and Evaluation

UX DESIGN FOR SPEECH INTERACTIONS

A*STAR
Institute for Infocomm Research

UX Design – An Introduction

What is USER EXPERIENCE ?

UX
Invisible
Unless
something
goes wrong …

Def. : *When a person is interacting with a product or a service has* ***an experience****; we call the person '****user****' and the experience, '****user experience****'.*

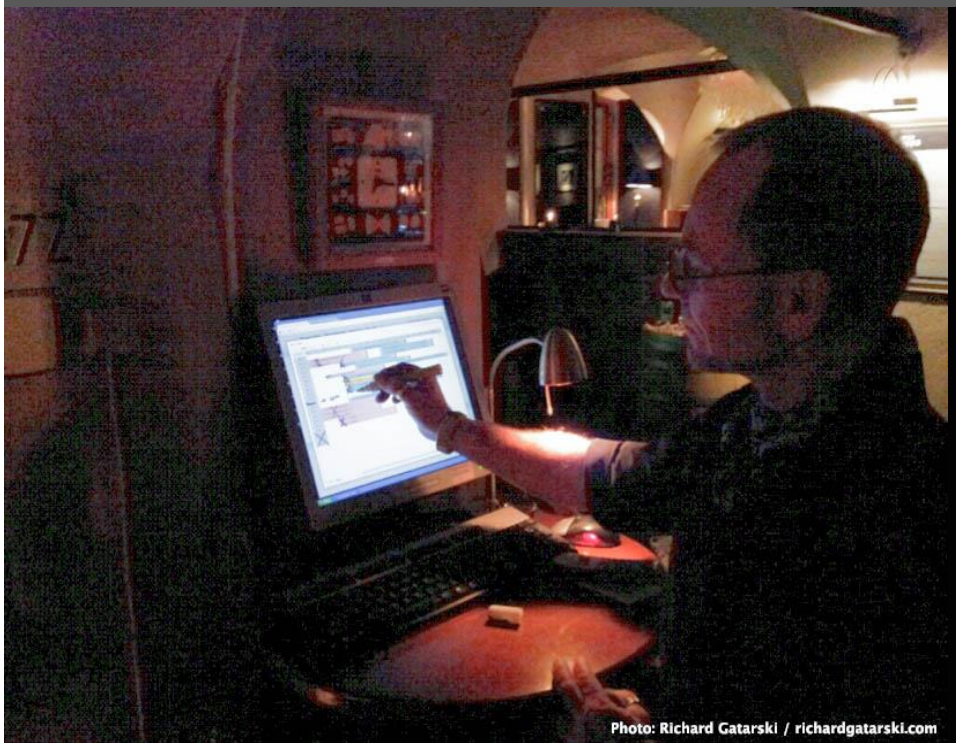
Basically, UX refers to the emotion, intuition and connection a person, aka a user feels when using a product or a service.

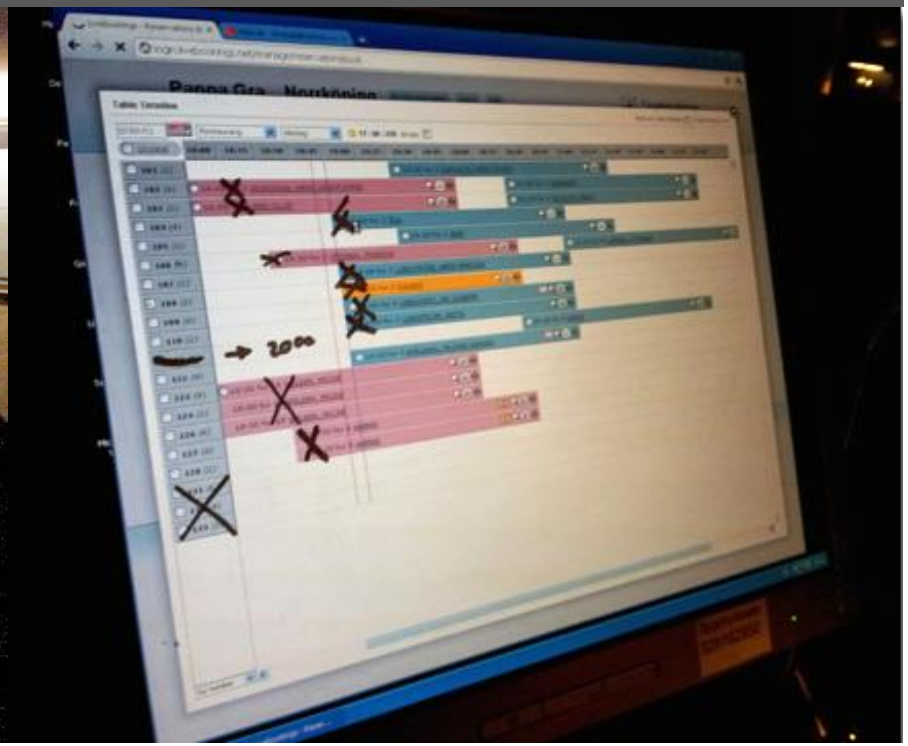
Main UX Design Principles

Match user expectation & context of use!

When UX Design Goes Wrong

Restaurant booking system*



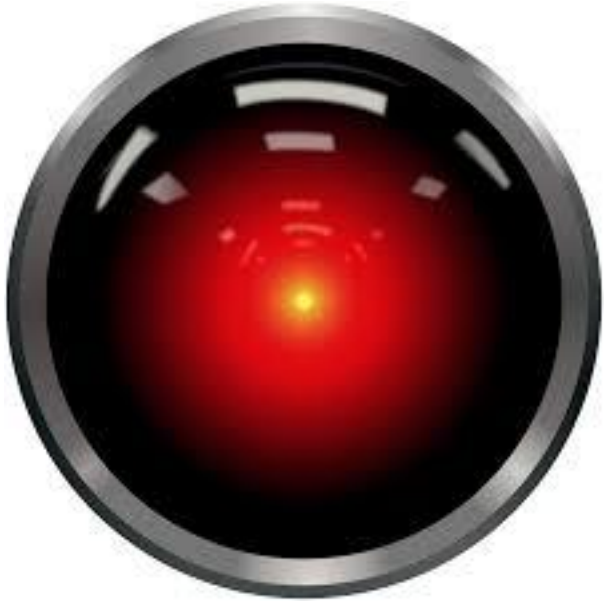


*D. Travis, User Experience (UX): The Ultimate Guide to Usability, Udemy academy 2013

User Experience vs. Design

Users will adapt the system to match their own way of work..

Designing Speech Interactions



"*Good afternoon, gentlemen. I am a HAL 9000 computer.*"
—HAL 9000 , 2001 A space Odyssey

Why Speech?*

- Speech is the most natural form of communication
- No practice is necessary
- Can be combined with other modalities
- It is fast

How fast … ?

Mode	CPM	Reliability	Practice	Others
Handwriting	200-500	Rec. error	No (requires literacy)	Hands & eyes busy
Typing	200-1000	~100% typos	Yes	Hands & eyes busy
Speech	1000-4000	Rec. error	No	Hands & eyes free

*C. Munteanu & G. Penn, Tutorial on Speech based interaction: Myths, Challenges & Opportunities, CHI 2015, Seoul, Korea

Designing Speech Interactions

However …

- Early hope for artificial intelligence have not been realized … Yet
- Communicating trough natural language is more difficult than anyone thought *
- Often frustrated by trivial errors when interacting with speech technology

* J.F. Sowa & Arun Majumdar –Kyndi Inc. "Natural language understanding". Data Analytics Summit 2015 II, Harrisburg, PA, Harrisburg University

DESIGNING SPEECH INTERACTIONS

Why all this is happening ?

Speech and language are highly complex processes

Computers

Are good with logic but aren't so good to relate, integrate & generalize concepts for all possible tasks & situations*

Humans

A 3 years old child is far better in learning & understanding speech than any currently available computer system*

Speech & Language

Both are highly contextual and contain an elevated degree of variability

*J.F. Sowa & Arun Majumdar –Kyndi Inc. "Natural language understanding". Data Analytics Summit 2015 II, Harrisburg, PA, Harrisburg University

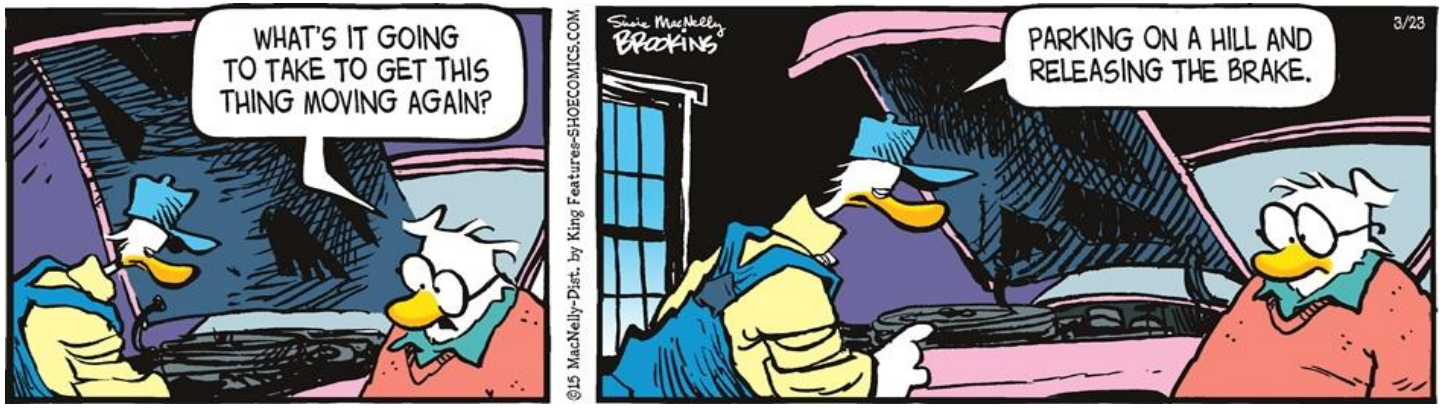
WHY IS SPEECH SO DIFFICULT TO PROCESS?

- **COMPLEXITY**
  - lots of data compared to text: typically 32000 bytes per second
  - Though classification problem: 50 phonemes, 5000 sounds, 100000 words
- **SEGMENTATION**
  - of phones, syllables, words, sentences
  - actually: no boundary markers, continuous flow of sample e.g. “I scream” vs. ”ice scream”, “recognize speech “ vs. “wreck a nice beach”
- **VARIABILITY**
  - acoustic channel: different microphones, different room, background noise
  - between speakers (different accents)
  - within speaker (e.g. respiratory illness)
- **AMBIGUITY**
  - homophones: “two” vs. “too”
  - similar sounding words: “wedding” vs “welding”
  - semantics: “crispy rice cereal” vs “crispy rice serial”

HARRY PICKED A BAD TIME TO GET LARYNGITIS

*C. Munteanu & G. Penn, Tutorial on Speech based interaction: Myths, Challenges & Opportunities, CHI 2015, Seoul, Korea

WHY LANGUAGE IS SO DIFFICULT TO PROCESS?



*http://www.shoecomics.com/

I. **Syntactic level:** understand the syntactic role of the words
   - parsing question & answer

RELATIVELY EASY

II. **Semantic level:** understand the meaning of the words
   - use background knowledge to understand:
     - Characters' roles & type of situation:
       "thing"= car ; "it" + "going" + "take" = finding a solution
     - physics law: a car on a slop, with no brakes

DIFFICULT

III. **Pragmatic level:** understand the meaning in context
   - understand irony & humor

VERY DIFFICULT

Speech recognition brings errors. Additionally, each level of language processing can introduce errors or ambiguities

*J.F. Sowa & Arun Majumdar –Kyndi Inc. "Natural language understanding". Data Analytics Summit 2015 II, Harrisburg, PA, Harrisburg University

What Is Required To Make It Work?

- Data, data and more data to train both ASR & NLU*
  - AM- ~ 100 hours of recorded speech in similar acoustic condition as the target domain + transcripts
    - Speaker dependent vs. speaker independent
    - Female/male
    - Read speech vs. conversational speech
  - LM – needs a large collection of texts similar to the target domain, e.g. Covering vocabulary, speaking style etc.
  - Classifier needs features vectors to be trained
- Use domain relevant data for training
- Restrict to a particular domain & input channel
- Do lots of adjustments

* C. Munteanu & G. Penn, Tutorial on Speech based interaction: Myths, Challenges & Opportunities, CHI 2015, Seoul, Korea

Other Critical Factors

- Digitization*
  - ❑ Sampling – ideally use a good sampling rate & don't change rates between recording and AM
  - ❑ Codecs – ideally uncompressed

- Microphone*
  - ❑ Best, if it has a fixed position, wind insulated, good sound-to-noise-ratio

- Latency
  - ❑ a big problem especially over slower networks

- Interaction design
  - ❑ dealing with user expectations, mainly determined by previous experiences

* C. Munteanu & G. Penn, Tutorial on Speech based interaction: Myths, Challenges & Opportunities, CHI 2015, Seoul, Korea

UX Design For Speech Interaction

Golden Rule:

1. Match user expectations & context of use

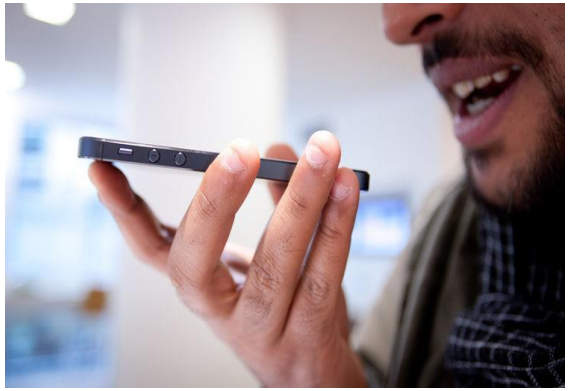
Anti Example: Speech in noisy environments

What people usually do when being in a noisy environment?

UX Design For Speech Interactions

Other Examples & Anti-Examples: Modality appropriateness

- **pointing** is more appropriate to determine a location on an interactive map*



- **speech** is more appropriate for filling "slots", like type of cuisine, departure time, dictation etc.*

- **mouse click** is more appropriate to open/close documents in a desktop environment*

Ideally, speech doesn't replace existing interaction, but rather enhance them**!**

**assuming users with no disabilities*

UX Design For Speech Interactions

We might never be able to get rid of ASR error but… To err is human!

2. Design for failure

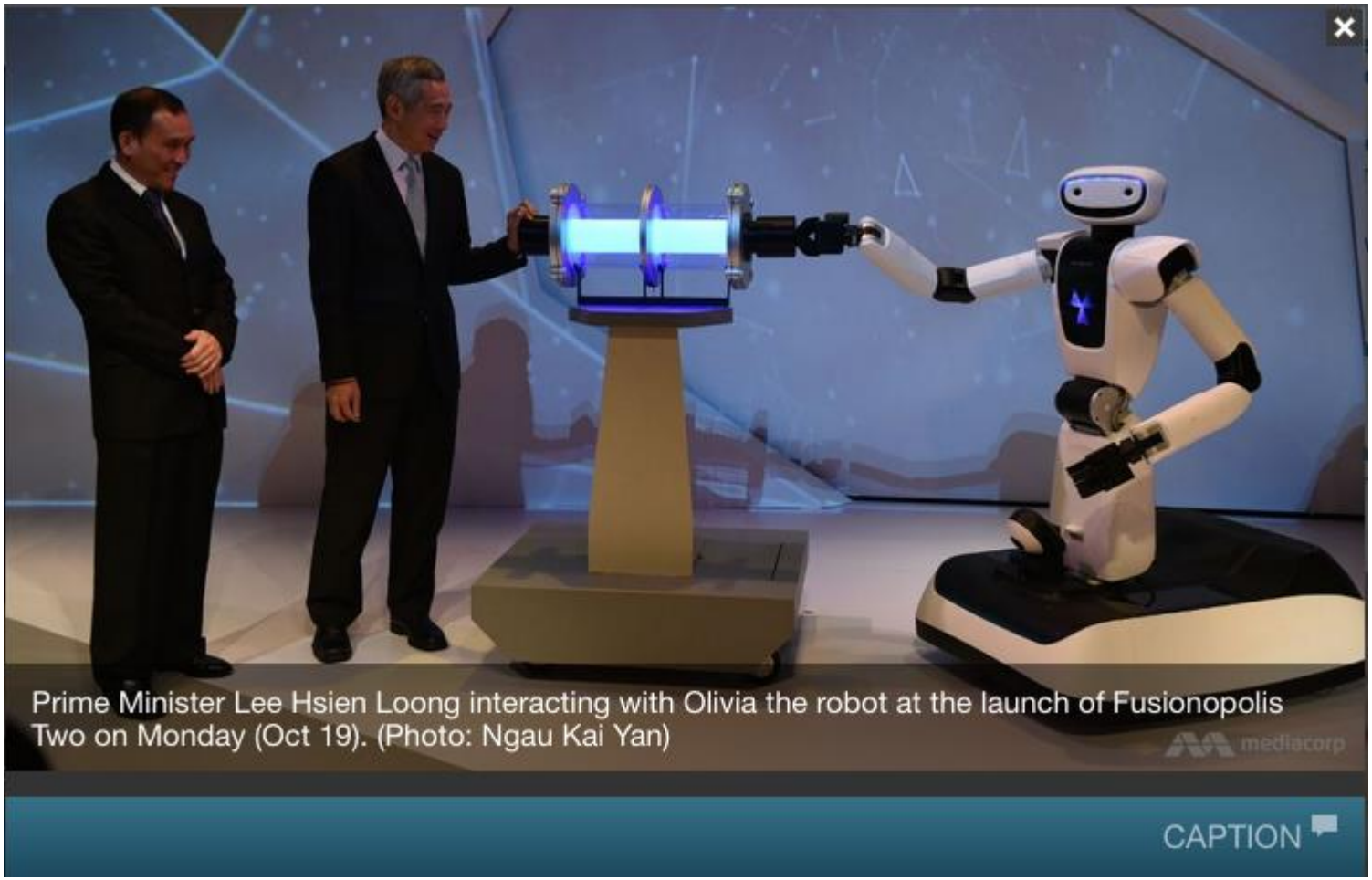
- Help users recover by offering input alternatives: multimodal interfaces offering touch input for typing
- Use complementary modalities, such as lips reading to enhance recognition performance*
- Use humor to overcome situations of failure and prompt users to repeat the input **
    - Do not irritated users!
    - Be aware of any cultural context (cultural usability)

* Petajan, E.D. Automatic lipreading to enhance speech recognition (speech reading), PhD Thesis 1984

*A.I. Niculescu, R.E. Banchs (2015) "Strategies to cope with errors in human-machine speech interactions: using chatbots as back-of mechanism for task-oriented dialogues", in Proc. of ERRARE 2015

Design For Failure

Example



Mechanical failure during rehearsal: Robot stepped back pretending being Overwhelmed by the crowd on the stage: “ I don’t like crowds” .

UX Design For Speech Interactions

3. Create personalization*

- Make the system appear smart
  Example: System: "*Last time you search for Thai. Would you like to search based on this type of cuisine again?*"

4. Ensure an appropriate expression manner**

- Avoid ambiguity, i.e. too open or non specific questions
  Choose instead to minimize the answer variability to a given question:
  Example: "Please specify your date of birth"
  Anti-example: "How can I help you?"
- Be short
- Announce breaks if the system needs time to process the information
- Choose an appropriate feedback type, e.g. **echo** or **implicit** feedback, i.e. that means embedding the feedback in the next system response
  Example (echo feedback):
  User: "*I am going to Copenhagen*"
  System: "*Copenhagen. What time would you like to depart?*"

* Johnny Schneider, Designing For Voice Interaction – UX Australia, Desining for Mobility , Melbourne Australia, 2013
**Niculescu, Andreea Ioana (2011) *Conversational* interfaces for *task-oriented* spoken dialogues: design aspects influencing *interaction* quality. PhD thesis.

Useful Application Cases

Don't obsess with using speech everywhere just because you can!

Question: Where to use it?

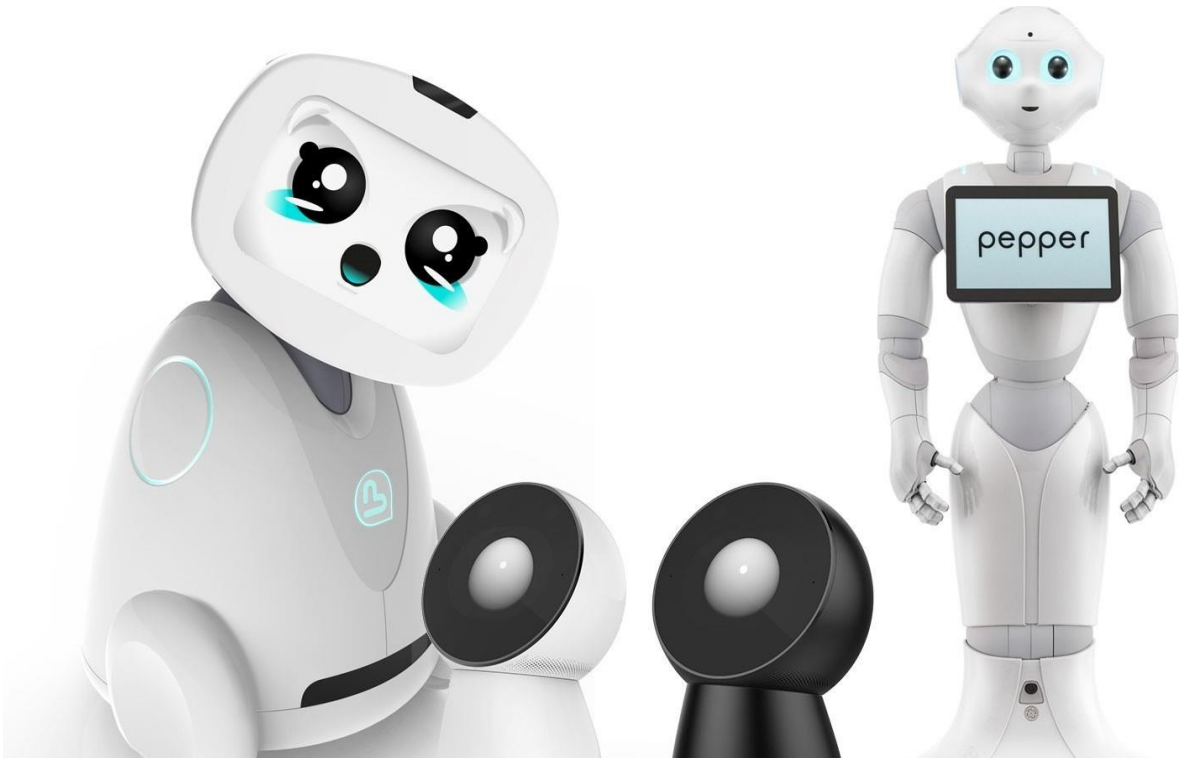
Answer: There were other modalities are not available or are not convenient to use. Ideally, speech is enhancing not replacing exiting interactions

Examples

- **Language tutorials**: learning pronunciation, vocabulary
- **Assistive technologies** for users with special needs (blind, motor impaired)
- **Gaming**: commands, faster than typing
- **Dictation tasks**: writing emails, faster than typing
- **Voice biometrics**: identifying someone without the subject's knowledge
- **Q&A engines for large amount of info:** recommendations, user manuals, exam questions
- **Social robotics:** makes interaction more natural

SPEECH IN HUMAN ROBOT INTERACTION (HRI)

Why speech for social robots ?



1. Convenient: Some robots are supposed to move around and perform tasks

2. No other input modality: No keyboard & no mouse

3. Comfortable: no need to Bend and input commands on the touch screen

4. More natural & human like: Speech is a human feature that brings HRI closer to human-human interaction

Additional design issues

1. Movement synchronicity:

- Embodiment gives more expression to HRI vs. simple speech interactions, but also requires synchronicity of **speech** & **gestures** & **behavior**: speech needs to match gestures, emotion expression and behavior

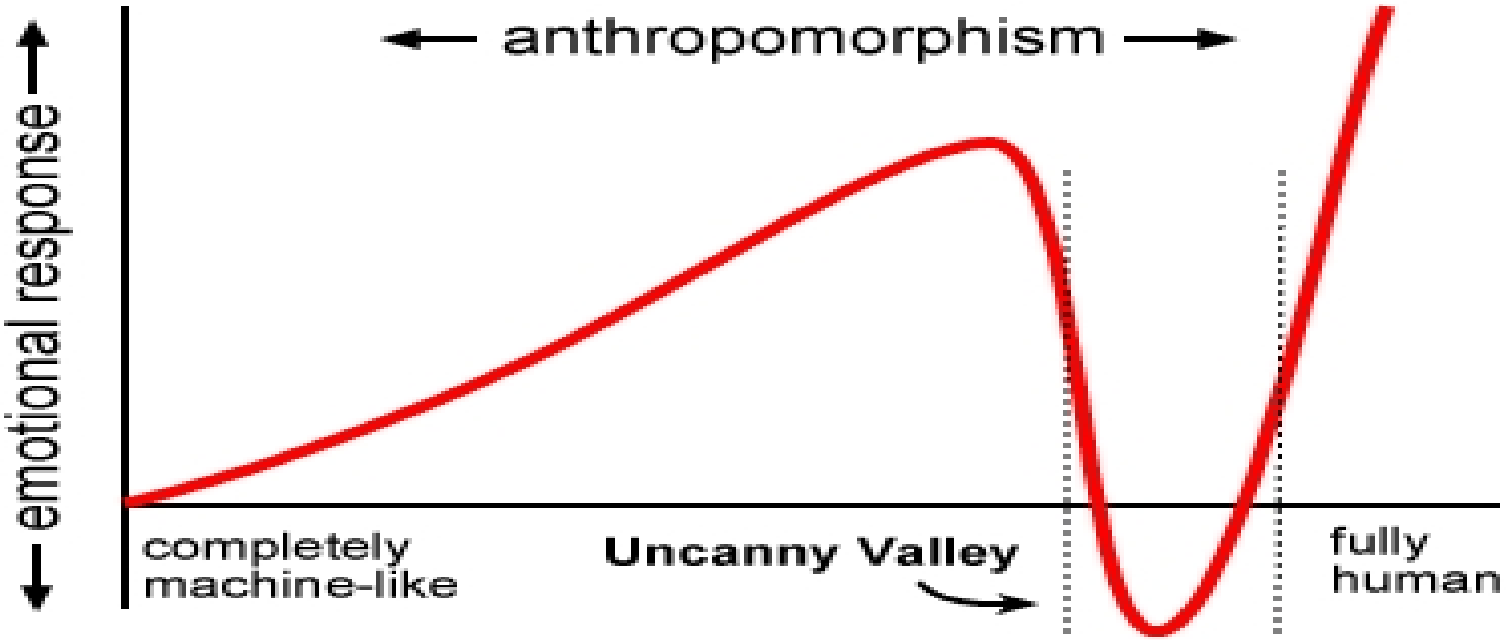
2. Appearance (voice & look) synchronicity:

- **Voice gender:** female vs. male recommended to match the robot appearance
- **Voice age**: children voice for a little robot

3. Personality synchronicity:

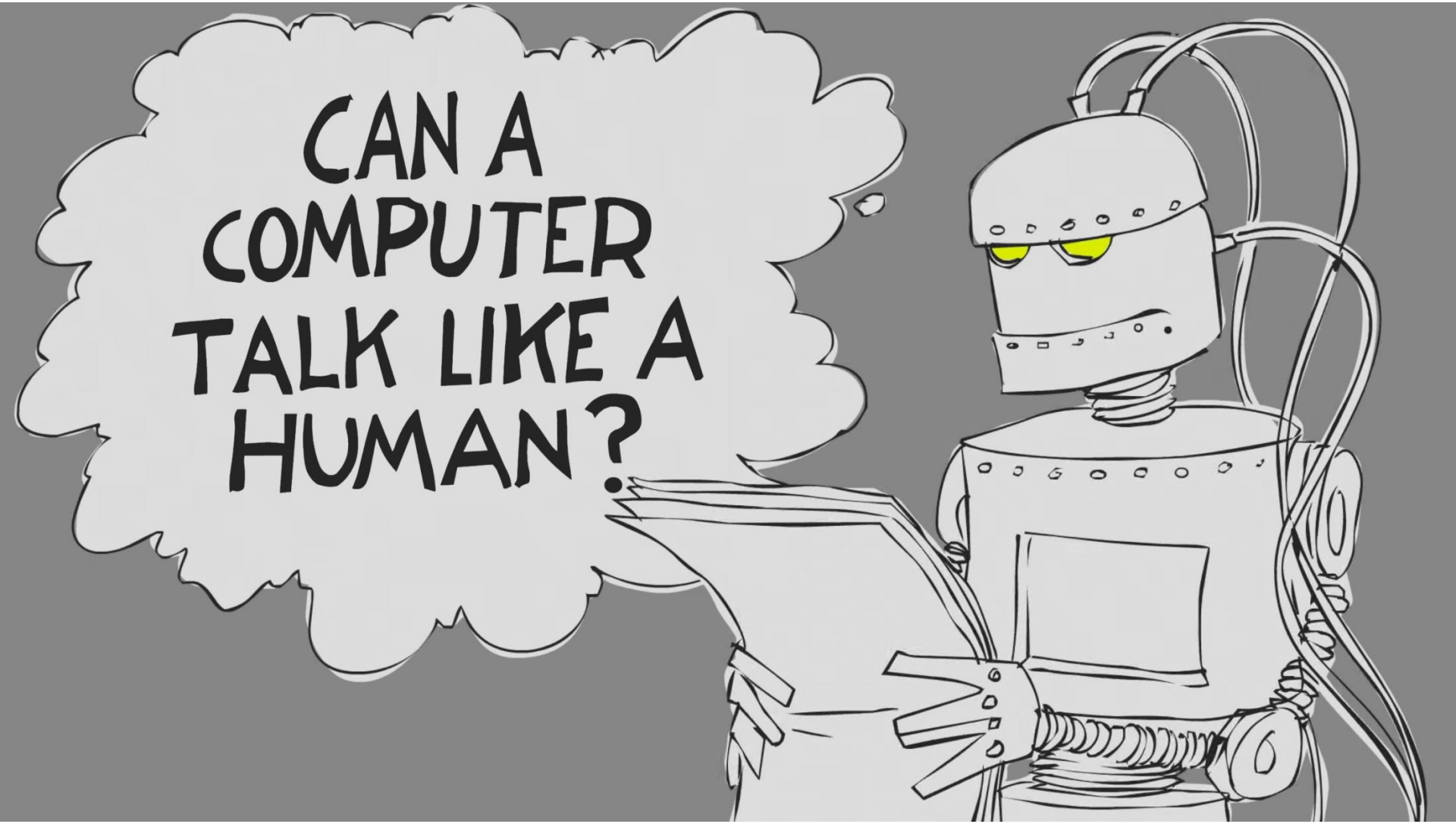
- **Voice tone, speech patterns** (framing), can all be used to express a certain personality type. Therefore, it is highly important to be in synchronicity with the **robot's behavior** and **movement**

UNCANNY VALLEY

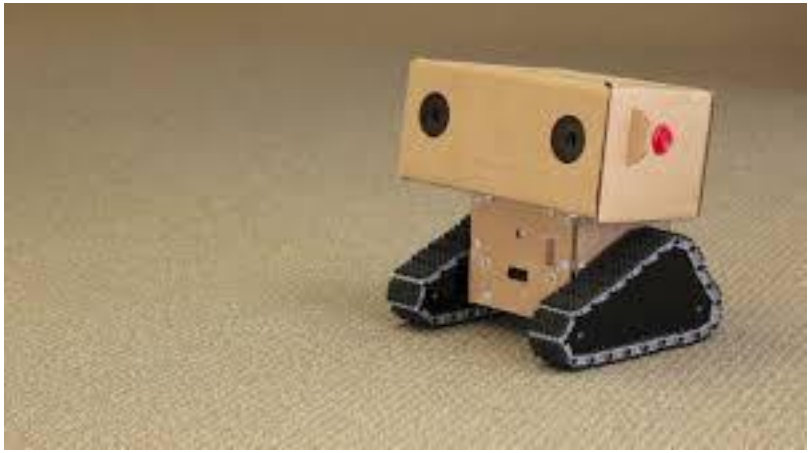


Still a long way to go …

SPEECH IN HUMAN ROBOT INTERACTION



Answer: NO!



Uncanny valley is not something we need to worry about … Yet ..

Part 4:
User Experience Design and Evaluation

USER STUDIES AND EVALUATION

Institute for Infocomm Research
A*STAR

USER STUDIES AND EVALUATION

Speech & **language** characteristics **(voice tone, voice accents, different formulation prompts etc.)** are amongst system features the easiest to control. Research studies have shown that manipulating them can change considerably the perspective users have about a speech interface

USER STUDIES AND EVALUATION

Study 1: Impact of English Regional Accents on User Acceptance of Voice User Interfaces

Research question: Would native Singaporean users prefer to speak with a virtual assistant speaking with a Singaporean accent as opposed to a British accent ?

Settings

- Controlled experiment with 59 users
- the users' task was to help users to easily find and use cell phone functions, such as SMS sending
- A voice talent recorded both Singaporean & British accented prompt sets
- Subjects were not told that the same person was playing the VA role in both cases.
- Questionnaires

Results

- Regardless of mother tongue, age, educational background or gender users tended to prefer the **British accent** over the Singaporean
- British accented voice was perceived as being **more polite**, $F(1, 58)=15.79$ $p<0.001$ & **having more sound quality**, $F(1, 58)=4.65$, $p<0.5$.
- Dialog with the British system was perceived as **being easier** than with its Singaporean counterpart

* A. Niculescu, G. M. White, S.L. See, R.U. Waloejo and Y. Kawaguchi (2008). Impact of English Regional Accents on User Acceptance of Voice User Interfaces. In Proc. of the 5th Nordic conference on Human-computer interaction, NordiCHI 2008, vol. 358, ACM, New York, pp. 523-526

USER STUDIES AND EVALUATION

Study 2: Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab*

Research question: What impact do have social skills have on the evaluation of a social robot?

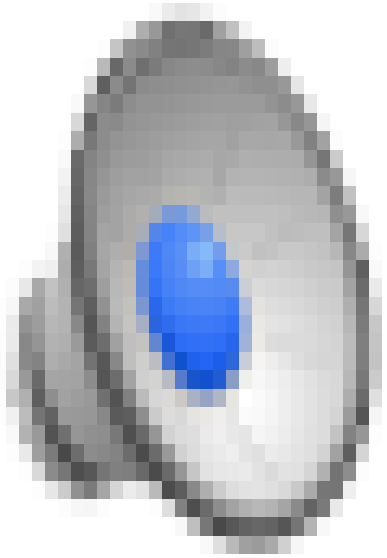
Settings

- Experiment in the wild with fully working prototype with 120 random people
- the users' task was to ask for information and play a simple game with the robot
- Questionnaire & Observations (Video recordings

*A.I. Niculescu, E.M.A.G. van Dijk, A. Nijholt, D.K. Limbu, S. L. See and A. H. Y. Wong (2010). Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab, in Proc. of the 2nd International Conference on Social Robotics, ICSR 2010, S.S. Ge, H. Li, J.-J. Cabibihan and Y.K. Tan (eds.), LNAI, vol. 6414, Springer Verlag, Berlin, pp. 50-62

User studies and evaluation

Study 2: Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab*



User studies and evaluation

Study 2: Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab

Results

- Ability to socialize was the second highest variable correlated with the overall interaction quality
- Robot's speech recognition performance was better ranked than the error logs predicted
- Generally, visitors were more tolerant to errors as compared to high latencies
- A pleasant voice is more important than a pleasant appearance

Final Remarks

- Speech/Voice and language are human characteristics that can trigger emotional responses in people

- People are "voice-activated": we respond to voice technologies as we respond to actual people and behave as we would in any social situation

- By taking this powerful finding, we can design voice interfaces can be user-friendly technology and achieve a better user acceptance since neither humans not technology is error–free.

*C. Nass and S. Brave. Wired for speech. How Voice Activates and Advances the Human-Computer Relationship. Cambridge MIT Press, USA, 2005.

CONCLUSIONS

- Human-Robot Interaction (HRI) is very complex since it is involves several types of problems: multidimensional (vison , language tactile etc.), multidisciplinary (Computer Science, Social Sciences, Psychology, Engineering etc.*)* and ill defined

- Speech and language technology – as part of the HRI) –is also very complex: natural language is ambiguous and contains lots of variability

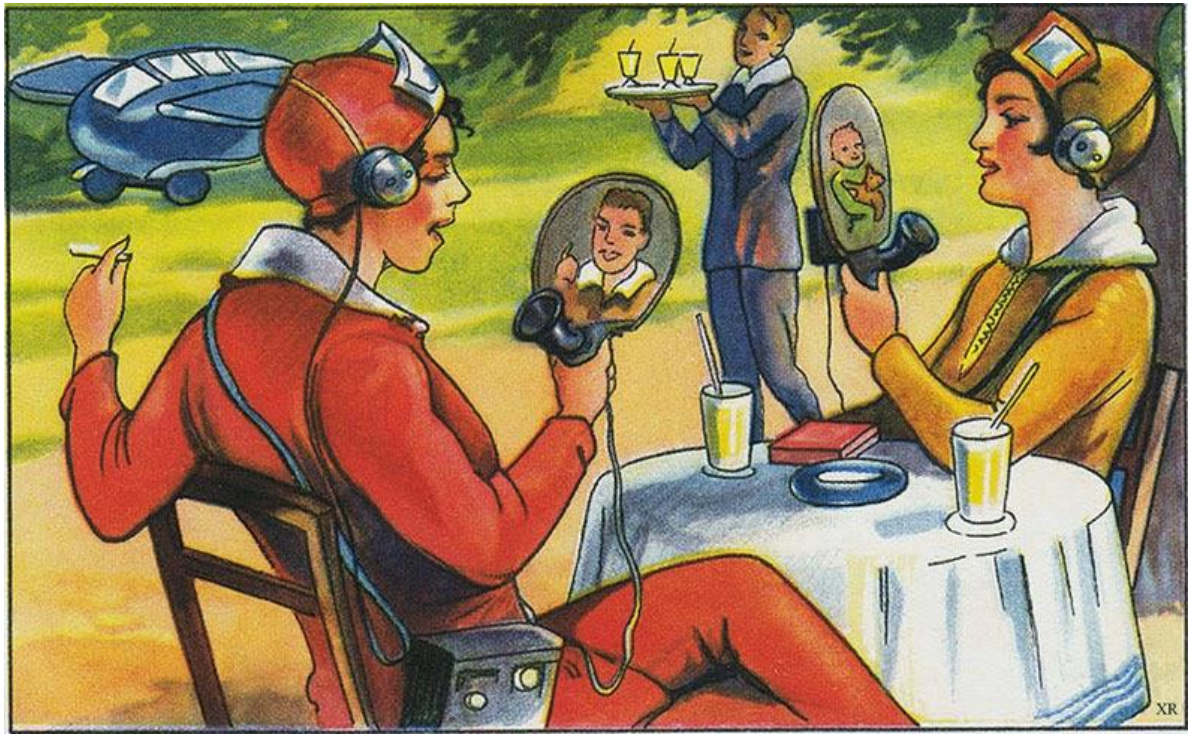
- Design plays a huge role in the user acceptance of technology- we just need to be aware of it

- Speech and language technology have a high potential for research and development, multiple possibilities still need to be discovered! Don't give up on it!

- Don't forget : Stay foolish and creative!

User studies & evaluation

"Whatever the mind can conceive and believe, it can achieve."
— Napoleon Hill, Think and Grow Rich: A Black Choice



This is our smartphone future, as imagined in 1930 by Hildebrands' Chocolate

☺

MAIN REFERENCES & ADDITIONAL REFERENCES

- Travis, D. User Experience- the ultimate guide to usability, Udemy academy 2013
- C. Munteanu & G. Penn, Tutorial on Speech based interaction: Myths, Challenges & Opportunities, CHI 2015, Seoul, Korea
- Sowa John F., Majumdar Kynd"Natural Language Understanding", Data Analytics Summit II (2015): http://www.jfsowa.com/talks/nlu.pdf
- N Almeida, S Silva, A Teixeira Design and development of a speech interaction: a methodology. International Conference on Human-Computer Interaction, 370-381 (2014)
- How to build up an NLU from scratch – a tutorial : http://www.vikparuchuri.com/blog/natural-language-processing-tutorial/
- Šabanović, S., Michalowski, M.P., Simmons, R. (2006). "Robots in the Wild: Observing Human-Robot Social Interaction Outside the Lab" *Proceedings of the IEEE International Workshop on Advanced Motion Control (AMC 2006),* Istanbul Turkey, March 2006.
- Johnny Schneider, Designing For Voice Interaction – UX Australia, Desining for Mobility , Melbourne Australia, 2013 http://www.slideshare.net/jonnyschneider/designing-for-voice-web
- Byron Reeves, Clifford Nass, The media equation: how people treat computers, television, and new media like real people and places. Center for the Study of Language and Inf, 2003
- Petajan, E.D. Automatic lip-reading to enhance speech recognition (speech reading), PhD Thesis 1984
- A.I. Niculescu, R.E. Banchs (2015) "Strategies to cope with errors in human-machine speech interactions: using chatbots as back-off mechanism for task-oriented dialogues", in Proceedings of Errors by Humans and Machines in multimedia, multimodal and multilingual data processing (ERRARE 2015)
- A. Niculescu, G. M. White, S.L. See, R.U. Waloejo and Y. Kawaguchi (2008). Impact of English Regional Accents on User Acceptance of Voice User Interfaces. In Proc. of the 5th Nordic conference on Human-computer interaction, NordiCHI 2008, vol. 358, ACM, New York, pp. 523-526
- C. Nass and S. Brave. Wired for speech. How Voice Activates and Advances the Human-Computer Relationship. Cambridge MIT Press, USA, 2005
- A.I. Niculescu, E.M.A.G. van Dijk, A. Nijholt, D.K. Limbu, S. L. See and A. H. Y. Wong (2010). Socializing with Olivia, the Youngest Robot Receptionist Outside the Lab, in Proc. of the 2nd International Conference on Social Robotics, ICSR 2010, S.S. Ge, H. Li, J.-J. Cabibihan and Y.K. Tan (eds.), LNAI, vol. 6414, Springer Verlag, Berlin, pp. 50-62
- A.I. Niculescu (2011) *Conversational* interfaces for *task-oriented* spoken dialogues: design aspects influencing *interaction* quality. PhD thesis

QUESTIONS?

Thank you for your attention!

A*STAR
Institute for Infocomm Research