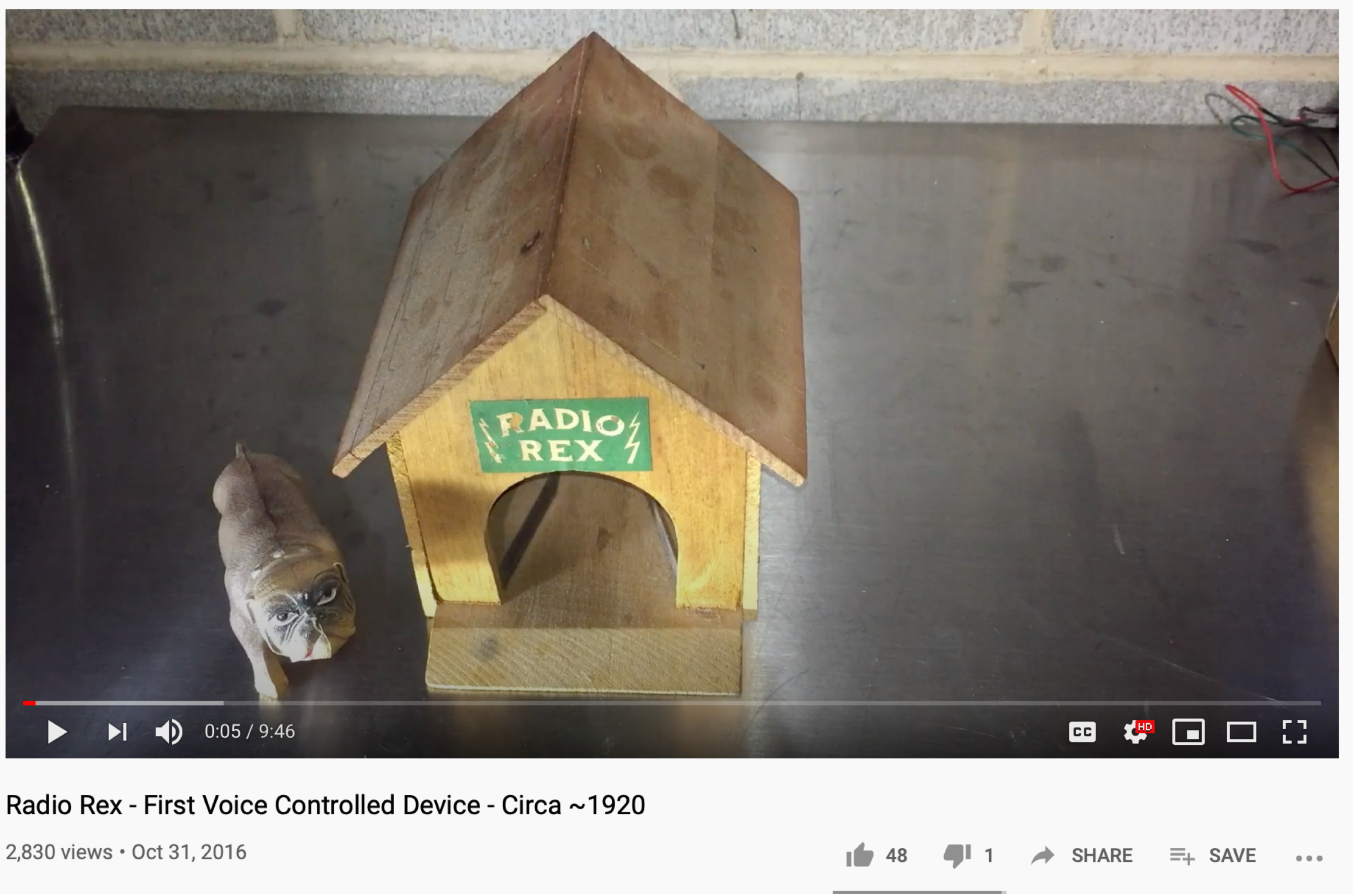
amazon alexa

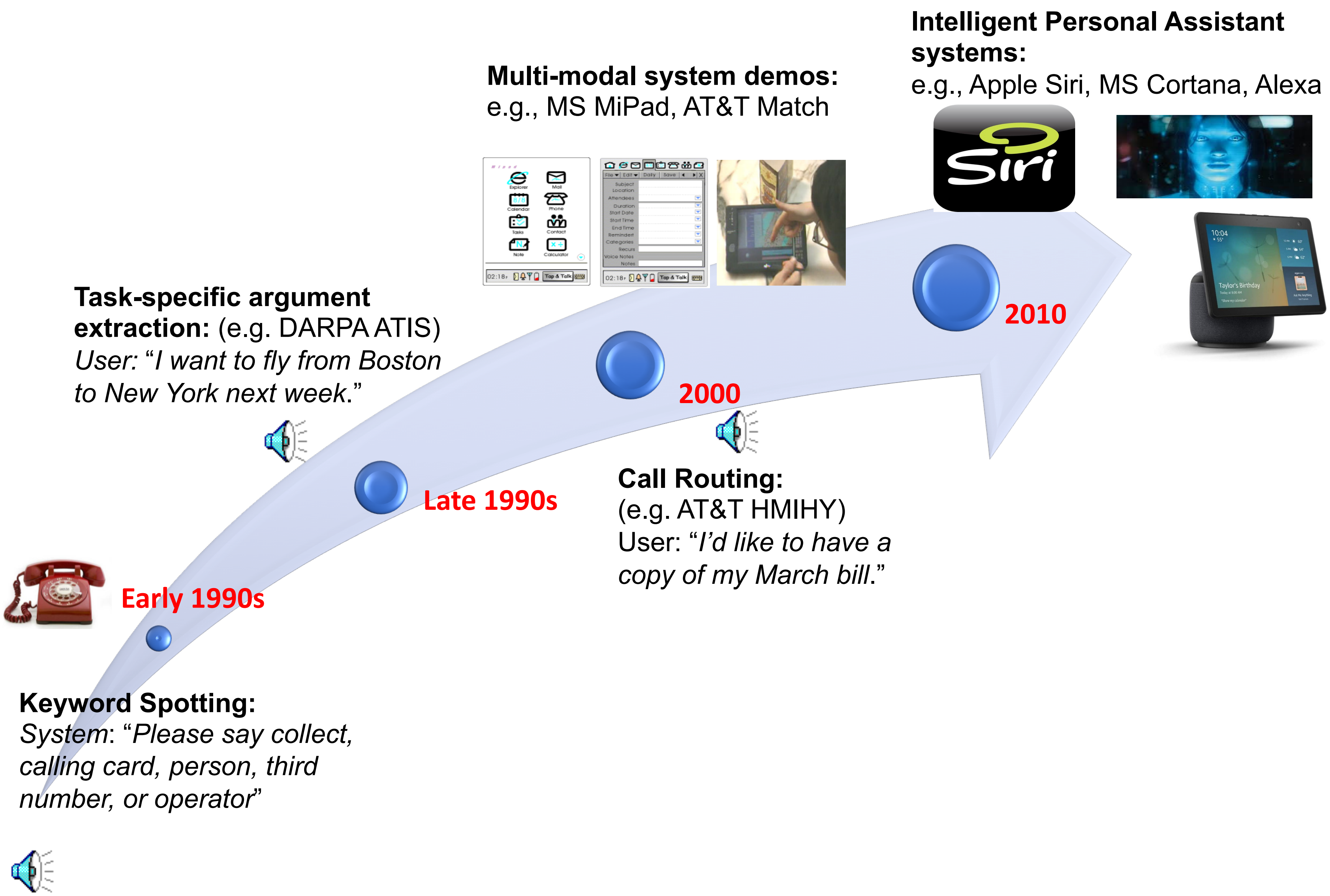
Interactive Learning for Conversational Understanding

Gokhan Tur
Alexa AI
December 2020

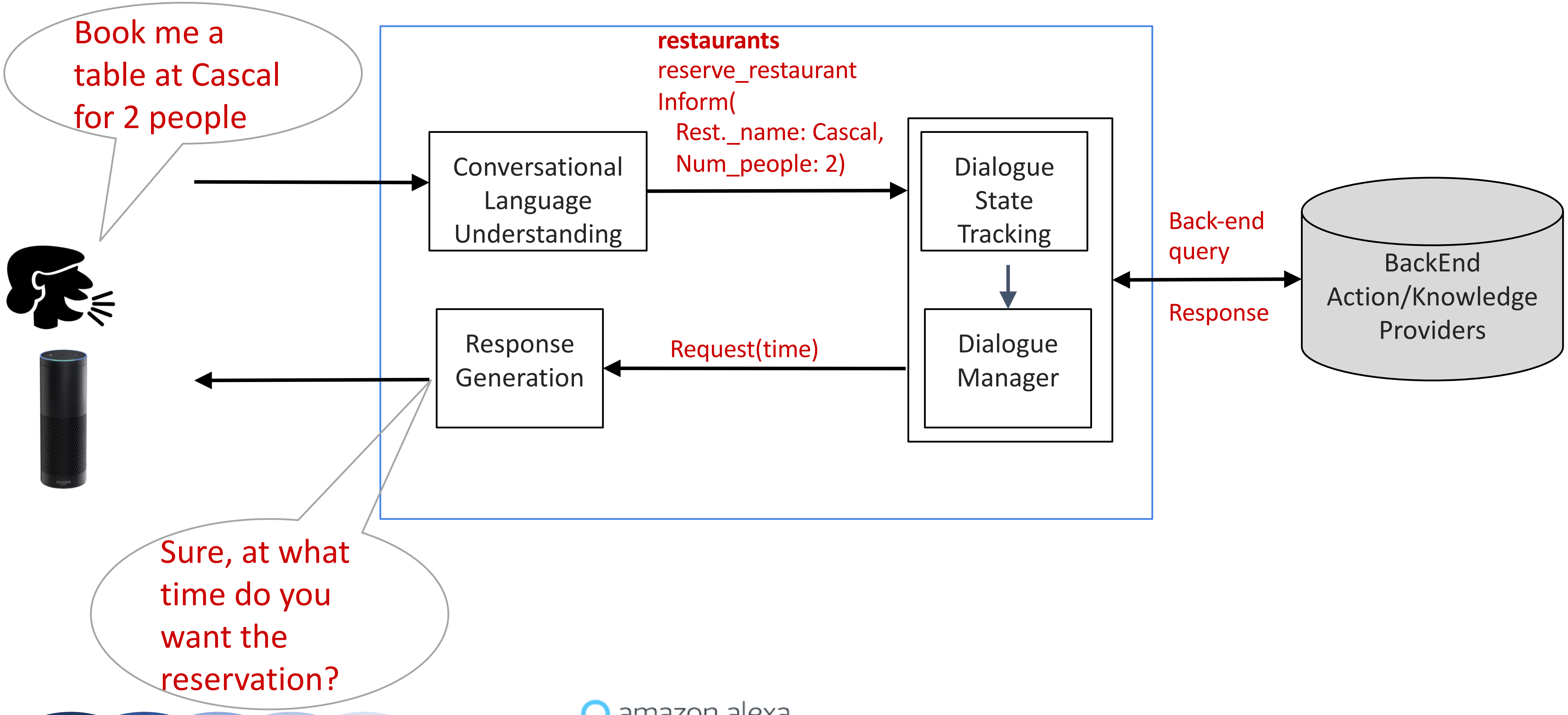
Dawn of Conversational AI



https://www.youtube.com/watch?v=AdUi_St-BdM


Intelligent Personal Assistant systems:
e.g., Apple Siri, MS Cortana, Alexa
Multi-modal system demos:
e.g., MS MiPad, AT&T Match
Task-specific argument extraction: (e.g. DARPA ATIS)
User: "I want to fly from Boston to New York next week."
2010
2000
Late 1990s
Call Routing:
(e.g. AT&T HMIHY)
User: "I'd like to have a copy of my March bill."
Early 1990s
Keyword Spotting:
System: "Please say collect, calling card, person, third number, or operator"

Task-Oriented Dialogue Systems (TODS)



Language Understanding in TODS

Semantic Representation: Flat or hierarchical frame of domain, intent, and slots.

DOMAIN = movies

"When was James Cameron's Avatar released?"

INTENT: Find_release_date
MOVIE NAME: avatar
DIRECTOR NAME: james cameron

Intents	Slots
Find movie	Movie genre
Find showtime	Movie award
Find theater	Theater location
Buy tickets	Number of tickets
…	…

DOMAIN = company

"Show me media companies in California"

INTENT: Find_company
LOCATION: california
INDUSTRY: media

Intents	Slots
Find company	Company name
Find revenue	Company address
Find founder	Company revenue
Find contact	Company industry
…	…

Domain/Intent Classification

- Mainly viewed as utterance classification.
- Given a collection of labeled utterances:

$$D= \{(u_1,c_1),\ldots,(u_n,c_n)\}$$

where $c_i \in C$, the goal is to estimate

$$c_k' = \operatorname{argmax}_{c \in C} P(c|u_k)$$

Example: "Show me the nearest movie theater"
Domain: movies
Intent: find-theater

Slot Filling

- Word sequence classification
- Given a collection tagged word sequences,

$$S=\{(w_1,t_1),\ldots,(w_n,t_n)\},$$

where $t_i = t_{i,1},\ldots,t_{i,|ui|}$, and $t_{i,m} \in M$, the goal is to estimate

$$t_k' = \text{argmax}_t\, P(t|w_k)$$

Example:

flights	from	Boston	to	New	York	today
O	O	B-city	O	B-city	I-city	O
O	O	B-dept	O	B-arrival	I-arrival	B-date

LU for Goal Oriented Dialogue Systems

- LU has been a hot R&D area since early 90s
- LU error rate has significantly reduced, especially after 2012 Deep Learning era
- A. M. Turing (1950) – *"Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."*

LU for Goal Oriented Dialogue Systems

- So why doesn't Conversational Understanding feel like a solved technology unlike Speech Recognition or Image Classification?
  - Because we do not *truly* understand, **we only act as if we understand**.
- Ray Jackendoff (2002) – *"Meaning" is the holy grail for linguistics and philosophy*
- Shannon 1948: "*Semantic aspects of communication are irrelevant to the engineering problem*."

5. First-order word approximation. Rather than continue with tetragram, . . . , n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

LU for Goal Oriented Dialogue Systems

- We only perform targeted understanding
- Buying a movie ticket is intent number XX in domain number YYY to the model.
- The system has not lived the experience of watching a movie in a theater buying a ticket unlike some humans. It has only *read* about it. There has been no situational grounding.

Persistent Areas of LU Research

- Issues:
  - Second turn
  - Variability in natural language
  - Long distance dependencies
  - Domain/Intent scaling
  - ASR noise
  - Model overfitting
  - Out-of-domain requests
  - New, dynamic, streaming events and entities
  - Uncovered in-domain requests
  - …

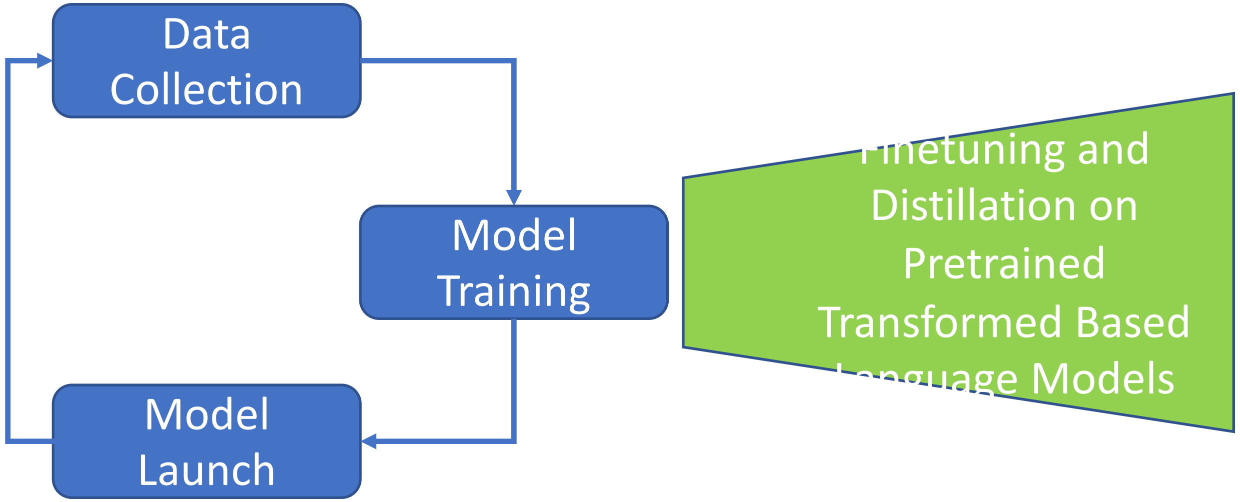
- Algorithms:
  - Contextual Modeling
  - Yet Another BERT based Model
  - Self Learning
  - Few Shot Learning
  - Joint ASR/LU Modeling
  - Multimodal Modeling
  - Offline intent/slot clustering
  - Teachable AI
  - …

Life Cycle of an LU system

- From Gary Marcus on GPT-3:

You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear ***the bathing suit to court. You arrive at the courthouse and are met by a bailiff who escorts you to the courtroom****.*

- But this does not mean that large transformer based pretrained language models are useless. On the contrary, the future of LU will be built on top of them.
  - … and probably on multimodal versions (e.g., ViLBERT or MAttNet)



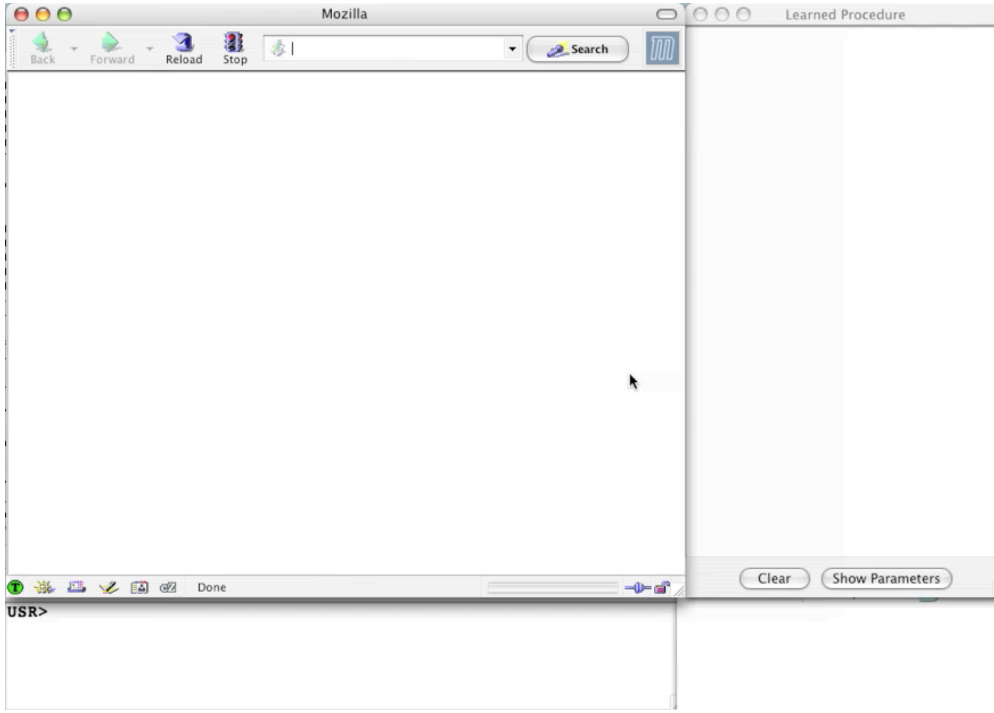
LU for Goal Oriented Dialogue Systems

- My vision: The only way we can solve ConvAI is to make it experience the real world, *interacting with users and things*, instead of offline supervised models trained for target domains.
  - Visual situational grounding for unknown objects:
    - *“look at my new porg plush”*

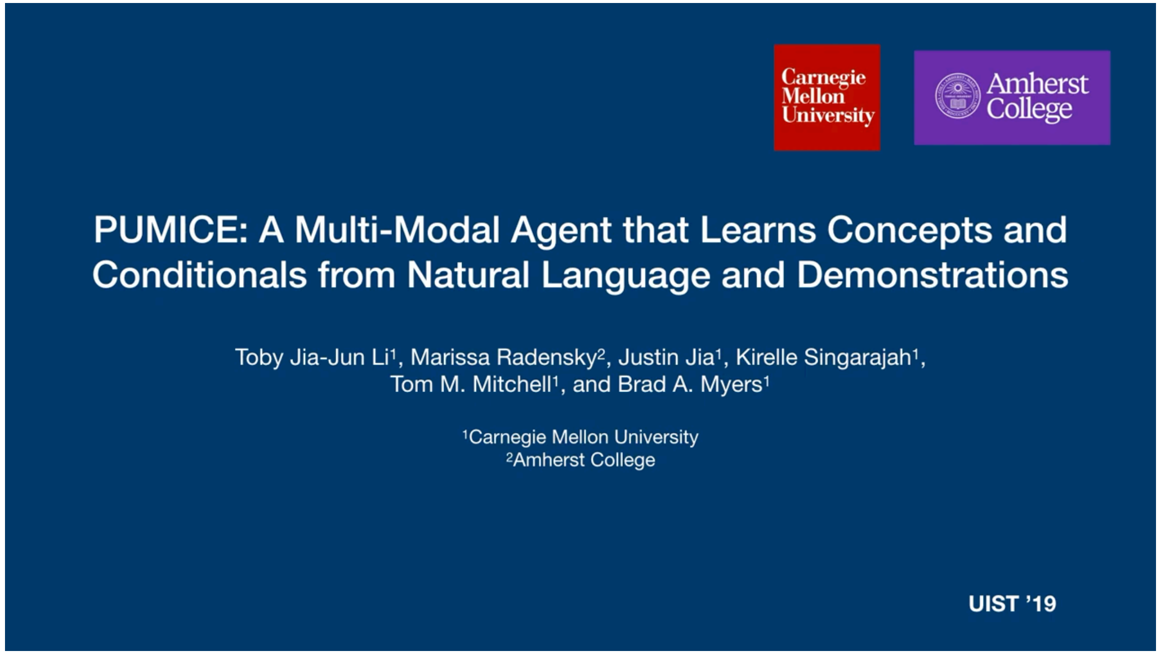
- Humans are very good at generalizing from few examples
- The conventional operation modes will change:
  - Supervised learning will not be mainstream.
  - Interactive self learning coupled with reinforcement learning will pave the way.

Interactive Learning

- Earliest study by Allan et al. 2007
  - *"Show me Gokhan Tur's papers"*
  - *"I don't know how to do that, can you teach me?"*

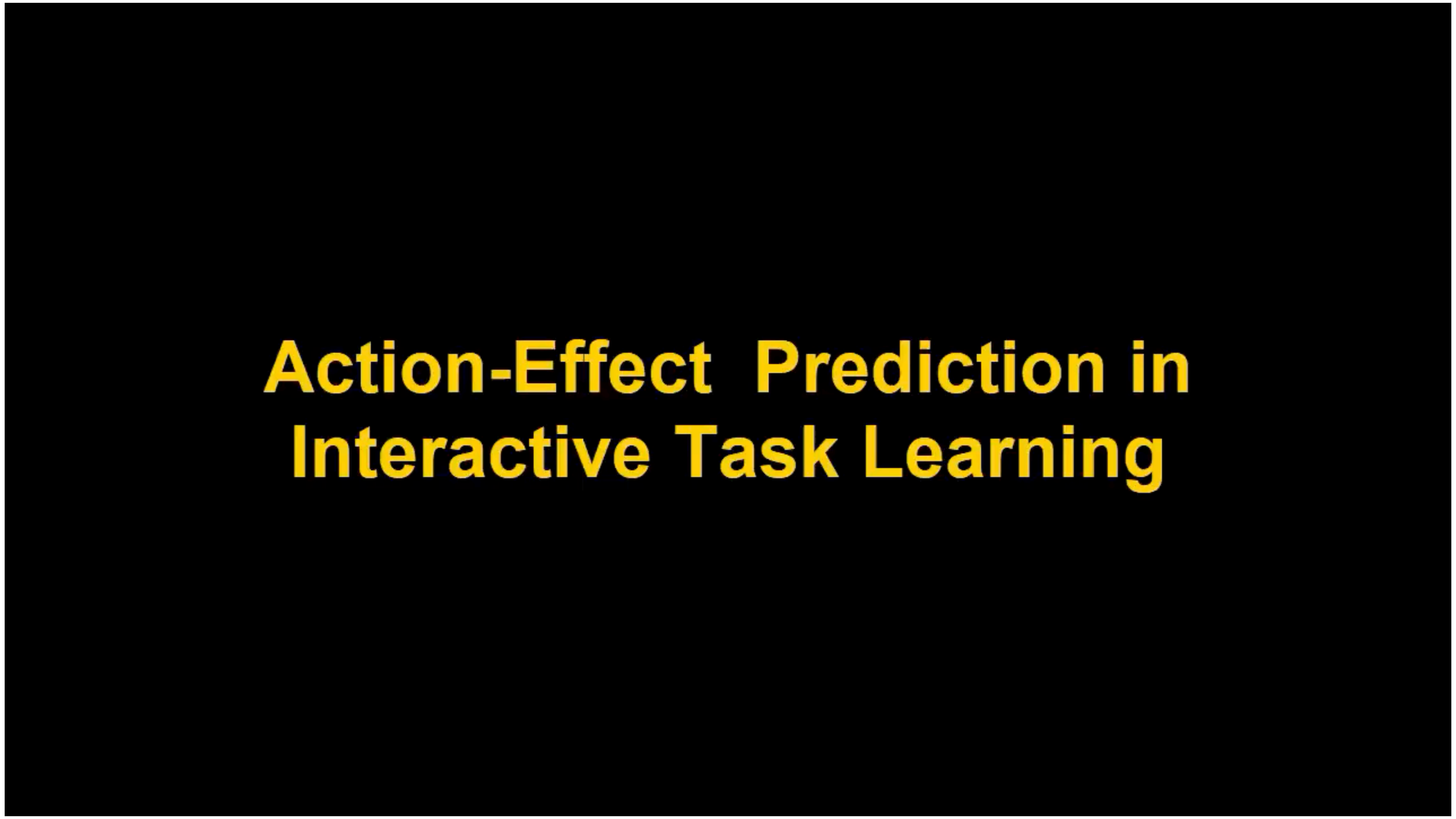


- Recent implementation by Li et al.



Interactive Learning

- HRI: Multimodal (Gao et al. 2017)



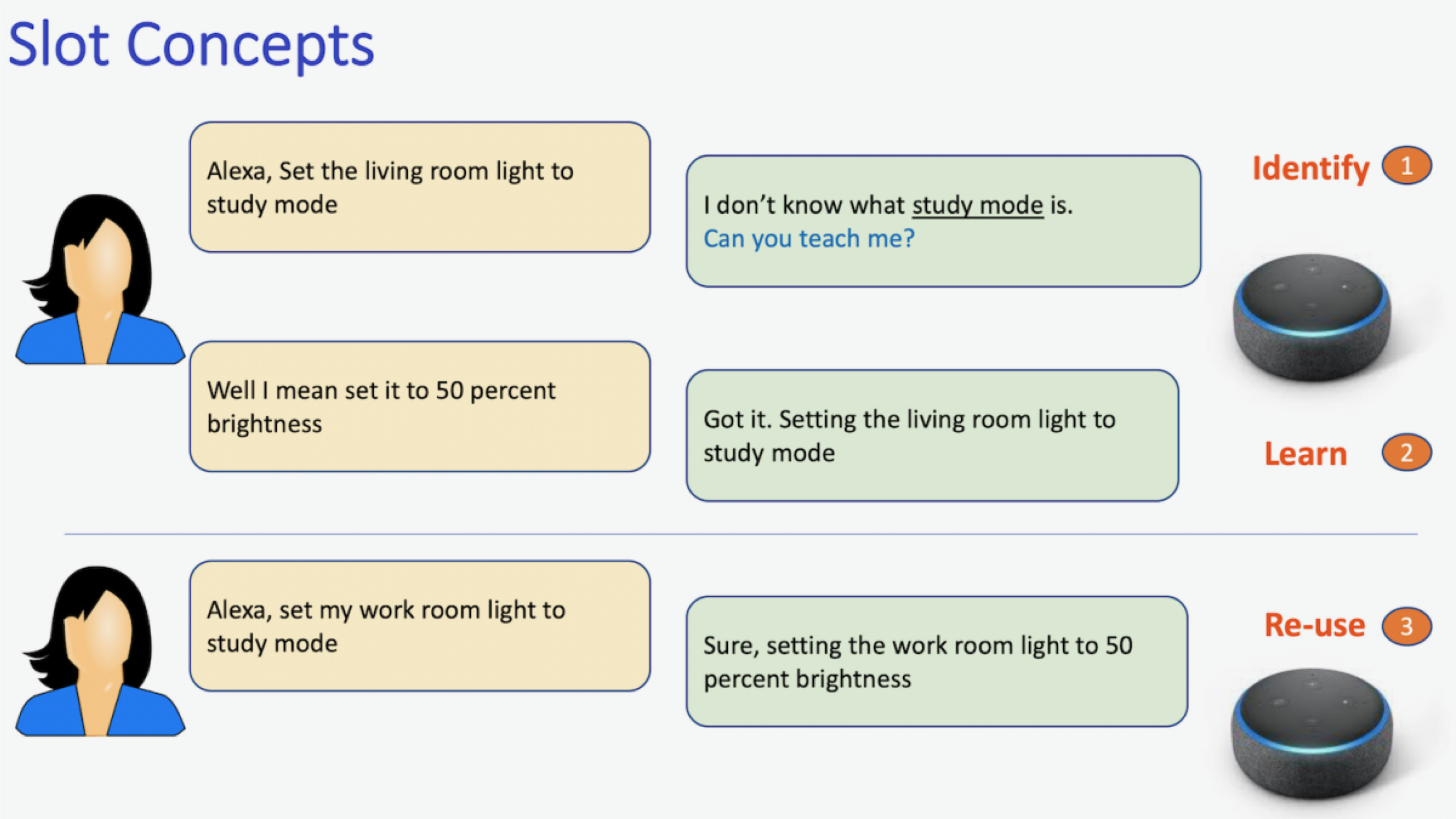
Interactive Learning

- 3 key research challenges:
  - How to model when to interact (“Can you teach me?”)
  - How to understand the response (“Let me teach you”)
  - How to reuse and generalize the learning

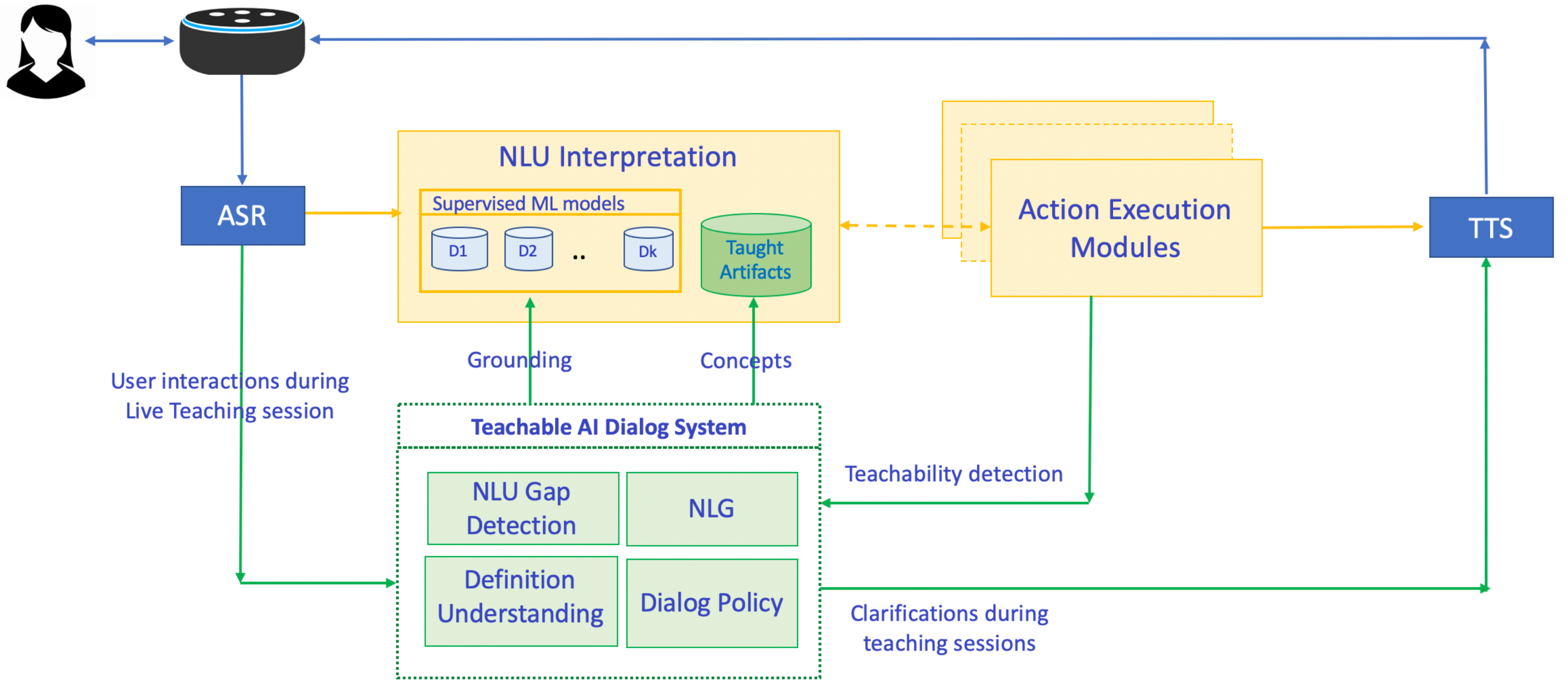
Concept Teaching

- Most relevant by Jia et al. 2017:
    - *“Buy a movie ticket for my birthday”*
    - *“When is your birthday?”*
- Amazon Alexa AI Interactive Concept Teaching paper in this session!

Concept Teaching



Concept Teaching



Concept Teaching – Model When to Interact

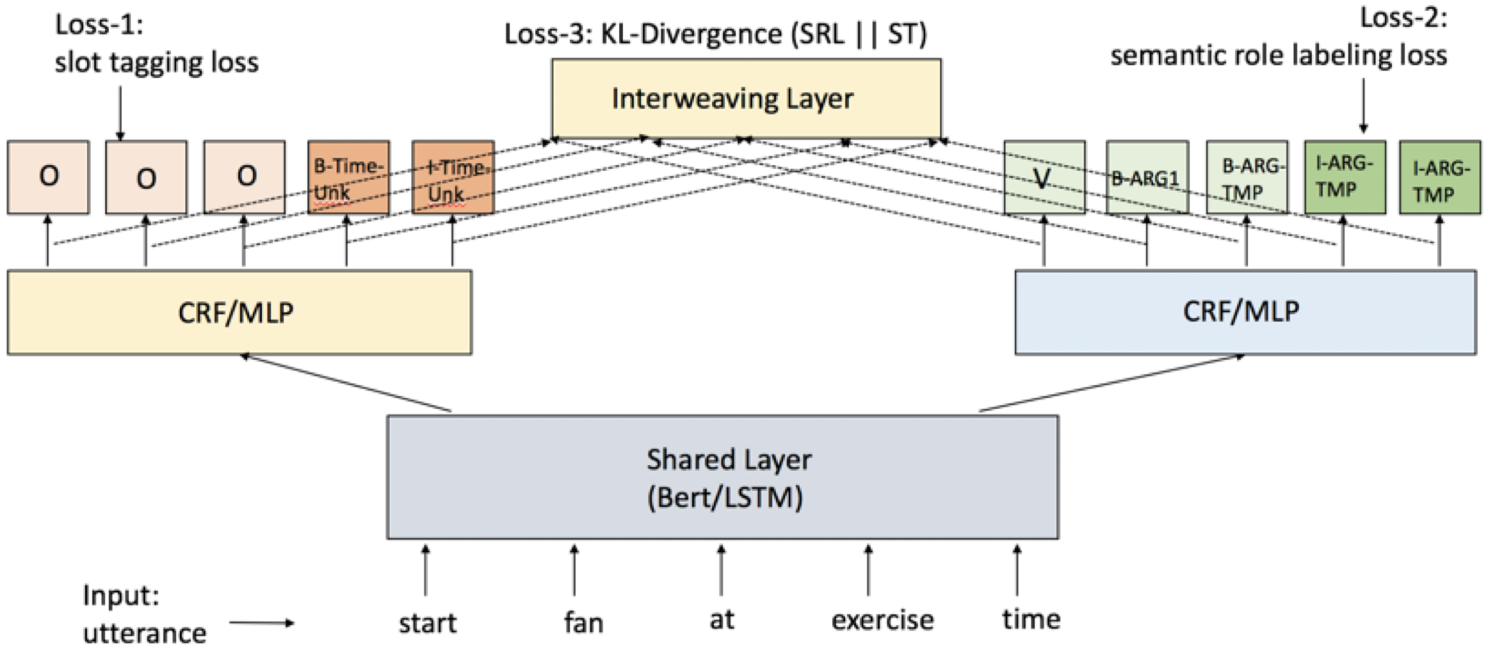
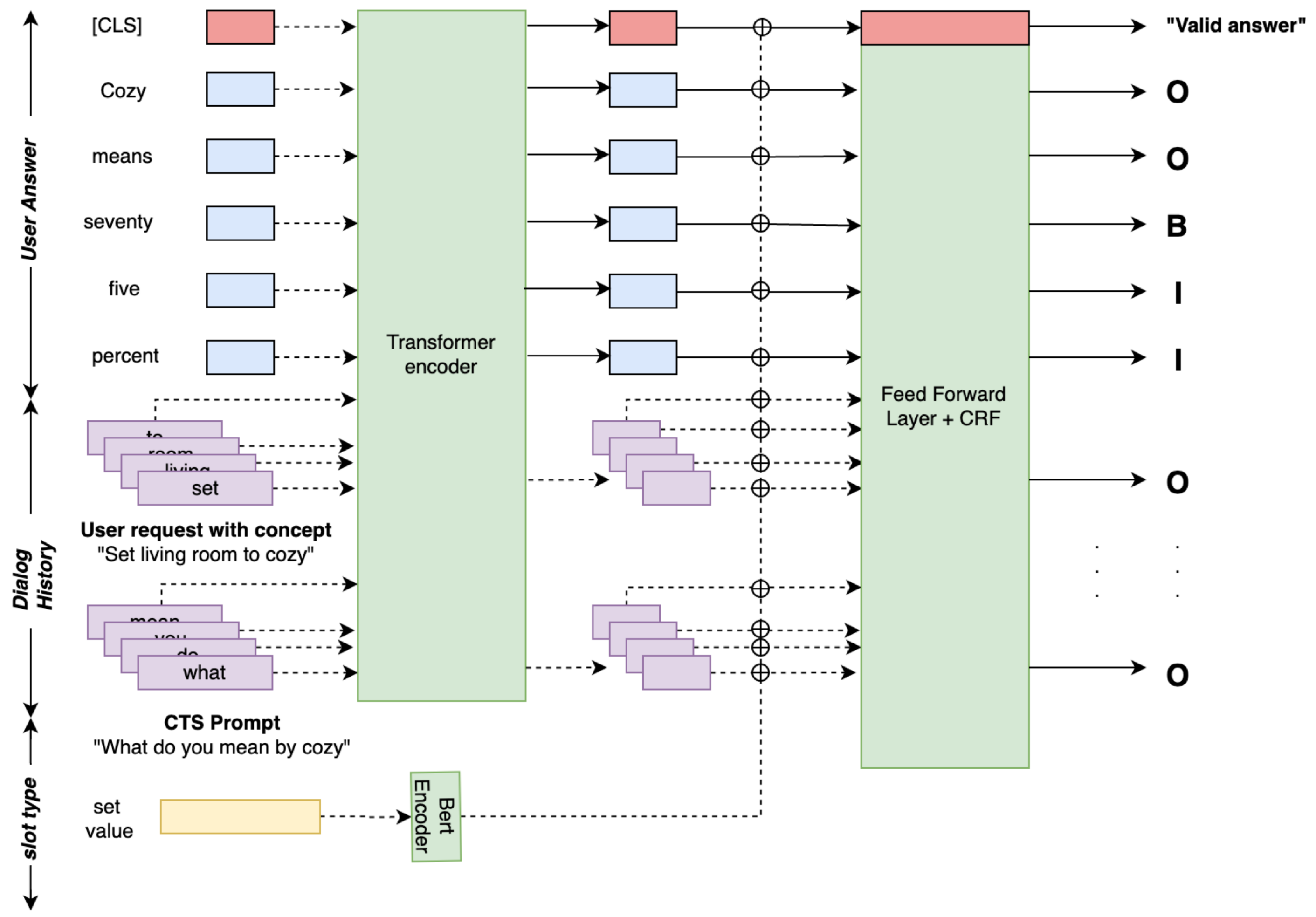


Figure 3. Multi-Task: Slot Tagging (ST) + Semantic Role Labeling (SRL) with Interweaving Layer

Concept Teaching – Understand the Teaching



HRI: Just Ask! (Chi et al. 2019)

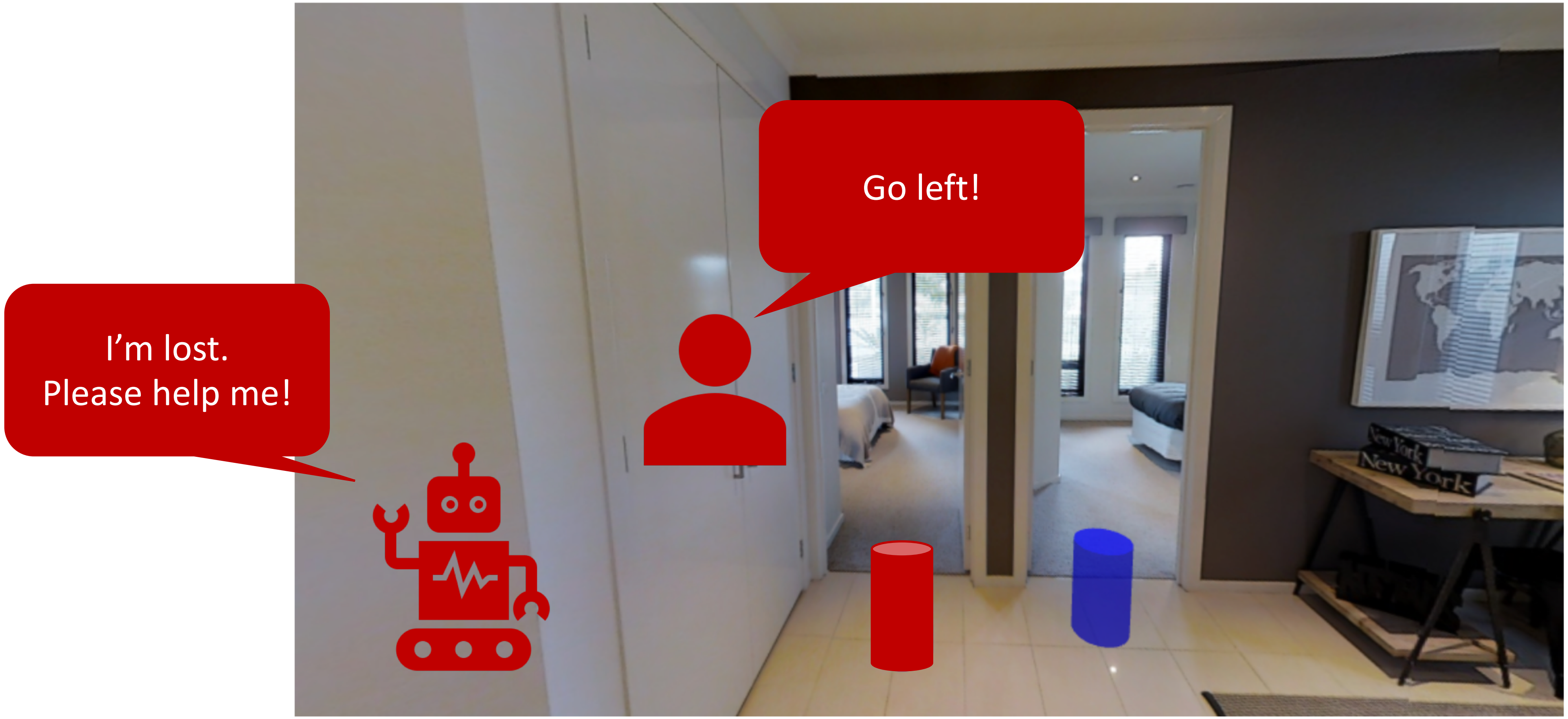
… Walk straight, right before you reach the bed.

HRI: Just Ask!

… Walk straight, right before you reach the bed.

HRI: Just Ask!

… Walk straight, right before you reach the bed.



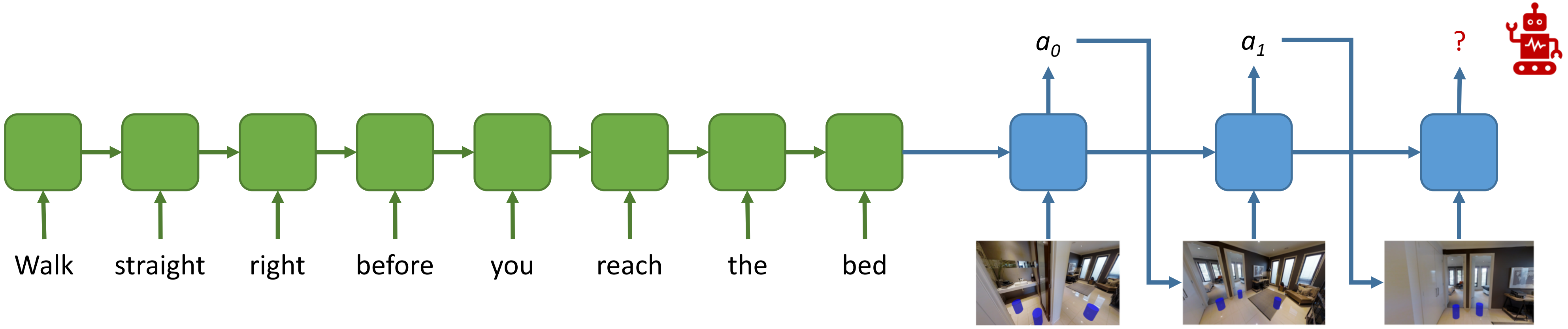
HRI: Just Ask!

… Walk straight, right before you reach the bed.

Thanks!

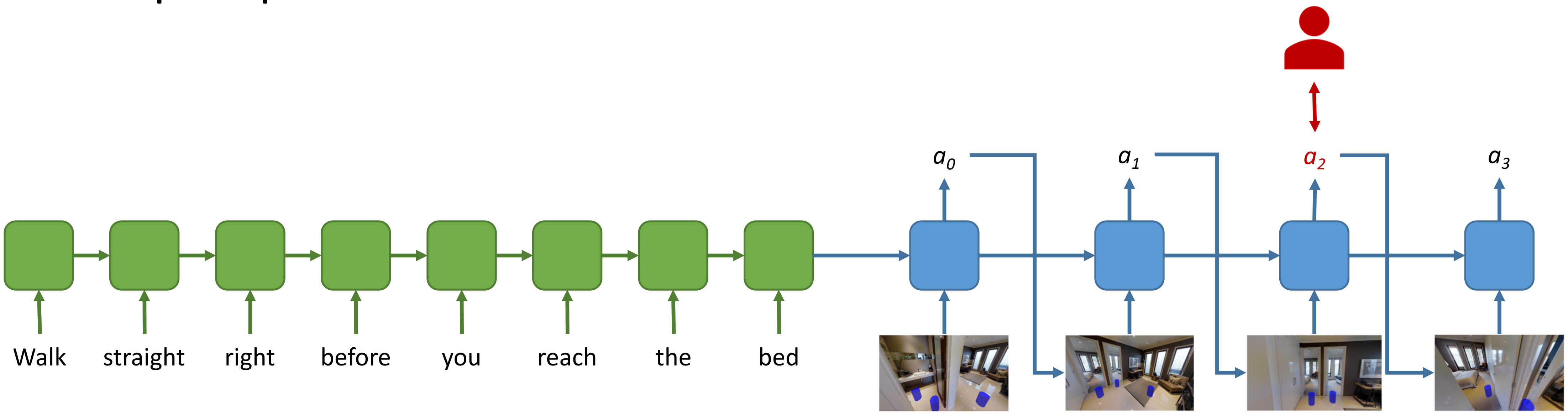
Proposed Model

- Seq2Seq with interactions



Proposed Model

- Seq2Seq with interactions



Summary

- Still scratching the surface on natural language conversational understanding after 30 years of research on goal oriented dialogue systems
- It is very possible that advances in computer vision will help language understanding significantly due to grounding.
- Personalization, reasoning, and active learning for interactive human-in-the-loop learning will be hot research directions for dialogue systems.