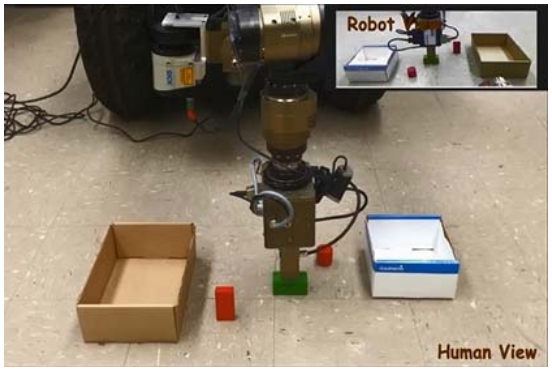
Language Grounding towards Situated Human-Robot Communication

Joyce Y. Chai

Language and Interaction Research Laboratory
Department of Computer Science and Engineering
Michigan State University

From SHRDLU to Human-Robot Communication

(Winograd 1972)



- Virtual or symbolic world: the world is known
- Actions are often deterministic
- Map language to internal symbolic representations

- Physical world: the world is unknown
- Actions may not be executed as expected
- Ground language to perception and action

Can traditional approaches be scaled up to enable human-robot communication

- Referential communication: interpretation and generation
- Action verb representation and interpretation

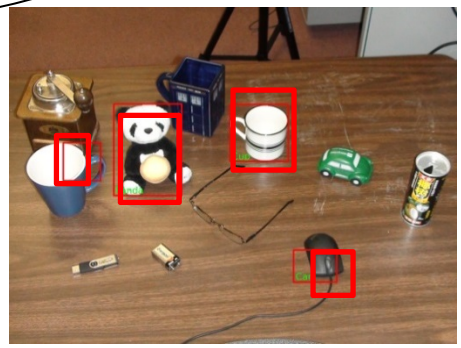
Referential Communication

Referring expression generation: Mimumum description with most distinguishable descriptors?

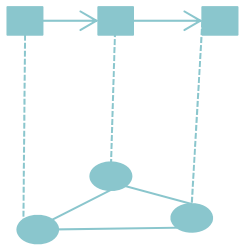
Reference resolution: entities that satisfy or best match the constraints?

"The small white object "

"the blue cup"

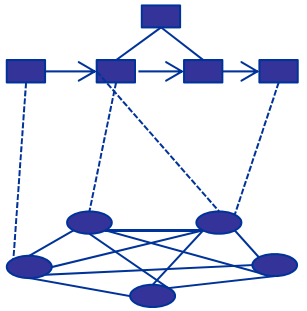


Robot's World Model



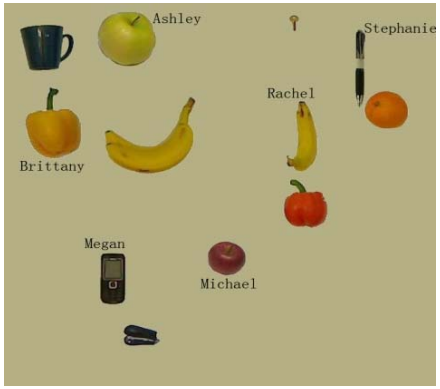
Mismatched perceptual basis

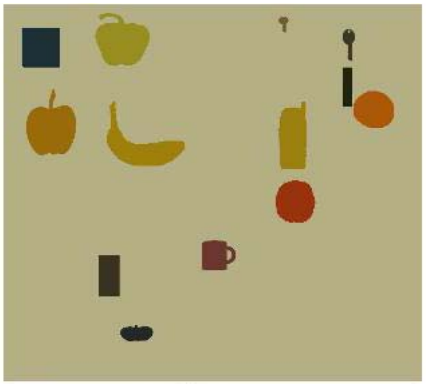
Human's World Model



How do people mediate perceptual discrepancies

Director





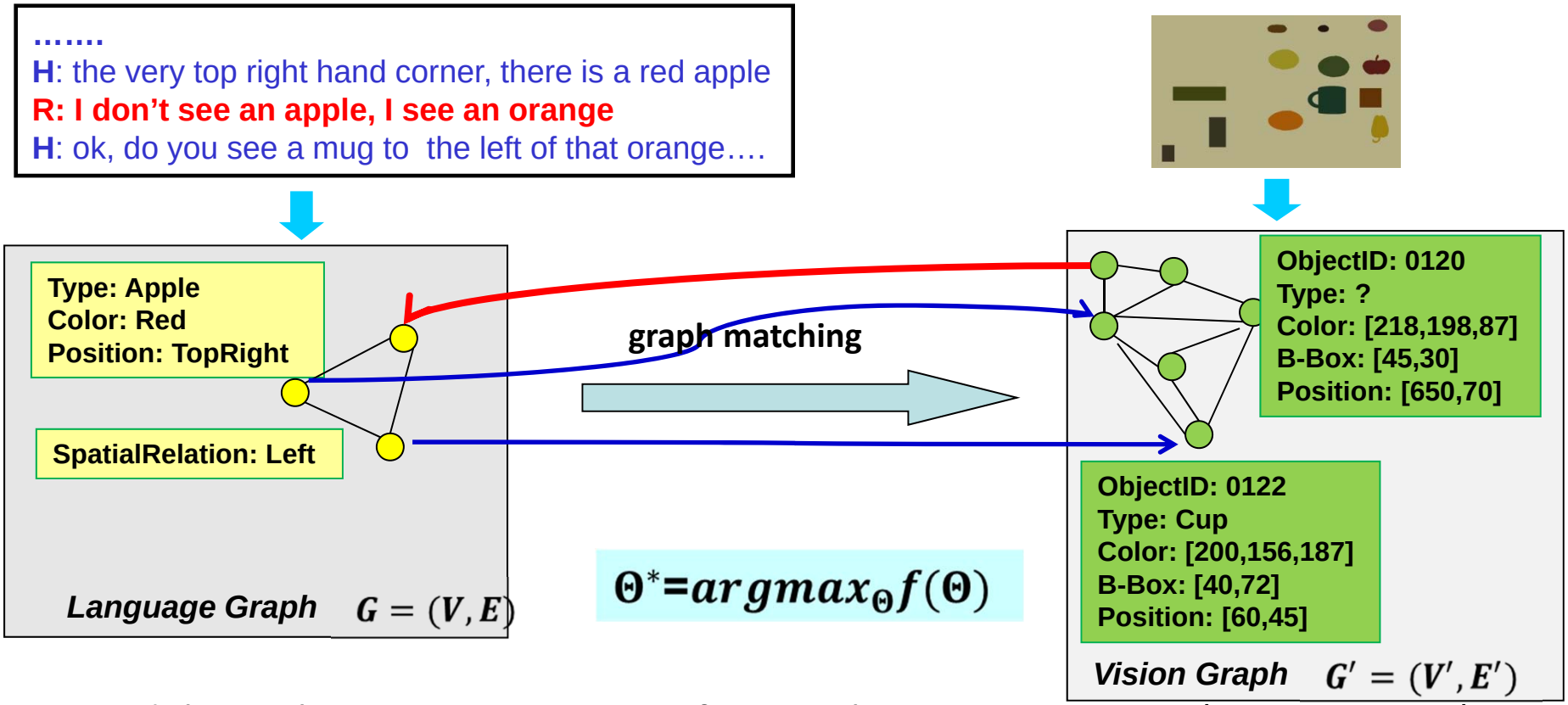
Matcher

D: there is basically a cluster of four objects in the upper left, do you see that?
M: yes
D: ok, so the one in the corner is a blue cup
M: I see there is a square, but fine, it is blue
D: alright, I will go with that, right under that is a yellow pepper
M: ok, I see apple but orangish yellow
D: ok, so that yellow pepper is named Brittany
M: ...

- Besides object-based properties, spatial relations and group-based spatial relations are commonly used
- Extra effort from the director (e.g., installment and trial)
- Extra effort from the matcher (e.g., proactive descriptions)

C. Liu, R. Fang, and J.Y. Chai. *Towards Mediating Shared Perceptual Basis in Situated Dialogue*. SIGDIAL 2012

Collaborative Reference Resolution

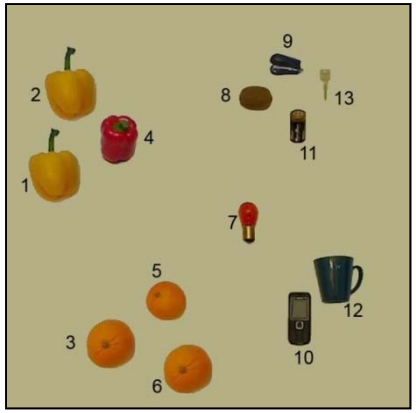


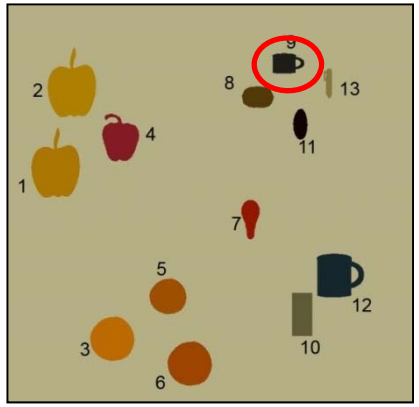
- Modeling relations compensates for visual processing errors (66% accuracy).
- Incorporating agent's collaborative behaivor reduces search space and improves grounding performance (18% absolute gain).
- Efficient algorithms are available to provide multiple hypotheses.

C. Liu, L. She, R. Fang, and J. Y. Chai. *Probabilistic Labeling for Efficient Referential Grounding based on Collaborative Discourse*. ACL 2014

C. Liu and J. Y. Chai. *Learning to Mediate Perceptual Differences in Situated Human-Robot Dialogue*. AAAI 2015

What about REG?





(Krahmer et al., 2003)

	Regular	Hypergraph	%gain
Matched perception (100% recognition performance)	80.4%	84.2%	**3.8%**
Mismatched perception (automatic recognition performance)	36.7%	45.2%	**8.5%**

Traditional competitive approaches fall apart when dealing with mismatched perception

R. Fang, C. Liu, L. She, and J. Y. Chai. *Towards Situated Dialogue: Revisiting Referring Expression Generation*. *EMNLP 2013*

Collaborative Models for REG

(Clark and Wilkes-Gibbs 1986) Referential communication is a collaborative process. To minimize the collaborative effort, partners tend to go beyond issuing an elementary referring expression, but rather using other different types of expressions such as episodic, installment, self expansion, etc.

- **Episodic**: two or more easily distinguished episodes

  A: below the orange, next to the apple, it's the red bulb.

- **Installment**: solicit feedback before moving on to the next

  A: under the pepper
  B: yes.
  A: there is a group of three objects.
  B: OK.
  A: there is an a yellow object on the right within the group.

Apply reinforcement learning to learn a policy for episodic/installment model

	Accuracy
Non-collaborative	47.2%
The episodic model	53.6%
The installment model	**68.9%**

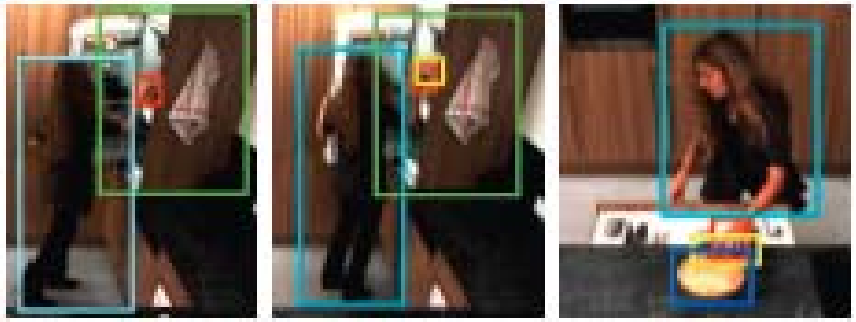
R. Fang, M. Doering, and J. Y. Chai. *Collaborative Models for Referring Expression Generation towards Situated Dialogue*. AAAI 2014.

Action Verb Representation and Interpretation

Predicate: take
- Arg0-PAG: Taker
- Arg1-PPT: thing taken
- Arg2-DIR: taken FROM, SOURCE of thing taken
- Arg3-GOL: destination

[The woman]$_{ARG0}$ [takes out]$_{Predicate}$ [a cucumber]$_{ARG1}$ [from the refrigerator]$_{ARG2}$.

The woman takes out a cucumber from the fridge.



(Regneri et al., 2013)

Predicate: take
- Arg0-PAG: track1
- Arg1-PPT: track2
- Arg2-DIR: track 3
- Arg3-GOL: track4

S. Yang, Q. Gao, C. Liu, C. Xiong, S. Zhu, and J. Y. Chai. *Grounded Semantic Role Labeling*, NAACL 2016

Physical Causality of Verbs

Linguistics studies have shown that concrete action verbs often denote some *change of state* as the result of an action (Hovav and Levin, 2010).

E.g., property of object, location, volume, area.

"He takes out a cutting board."

"She cuts the cucumber."

- Can we explicitly model physical causality?
- Can causality modeling of actions verbs provide high-level guidance for underlying vision processing and thus grounding language to the visual world?

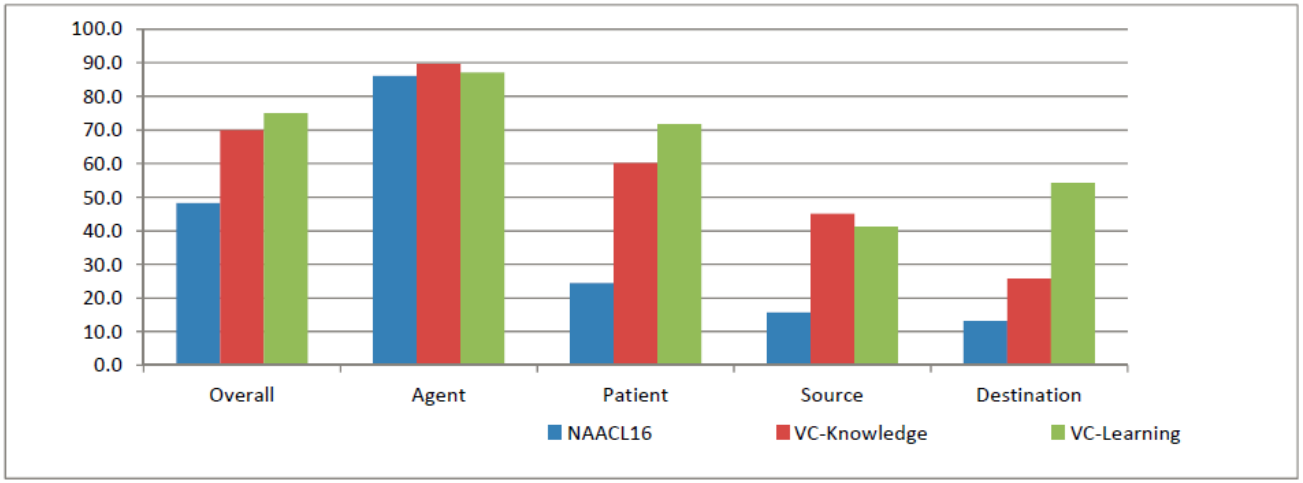
Causality of Verbs for Grounding Action Frames

Causality Knowledge

take: <location…>
cut: <size, quantity…>
wash: <wetness, color…>
.....

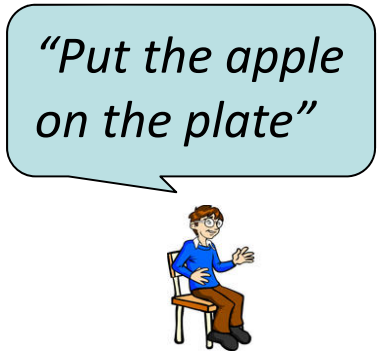
Visual Detector

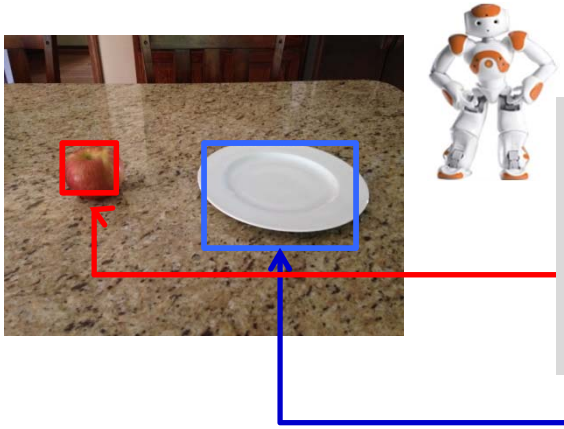
Attribute	Rule-based Detector	Refined Rule-based Detector
Attachment / NumberOfPieces	Multiple object tracks merge into one, or one object track breaks into multiple.	Multiple tracks merge into one.
		One track breaks into multiple.
Presence / Visibility	Object track appears or disappears.	Object track appears.
		Object track disappears.
Location	Object's final location is different from the initial location.	Location shifts upwards.
		Location shifts downwards.
		Location shifts rightwards.
		Location shifts leftwards.
Size	Object's x-axis length or y-axis length is different from the initial values.	Object's x-axis length increases.
		Object's x-axis length decreases.
		Object's y-axis length increases.
		Object's y-axis length decreases.



Q. Gao, M. Doering, S. Yang, and J. Y. Chai. *Physical Causality of Action Verbs in Grounded Language Understanding*. ACL 2016

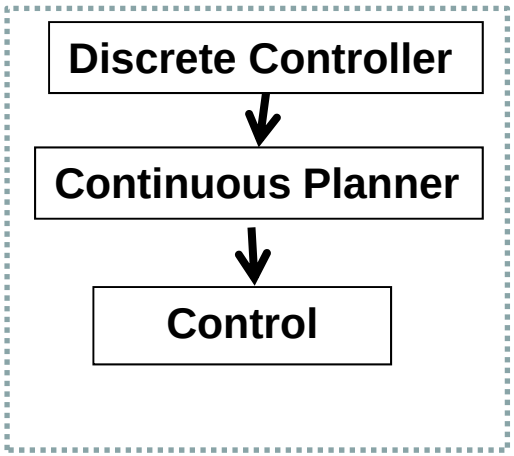
Is Grounded Action Frame Sufficient ?

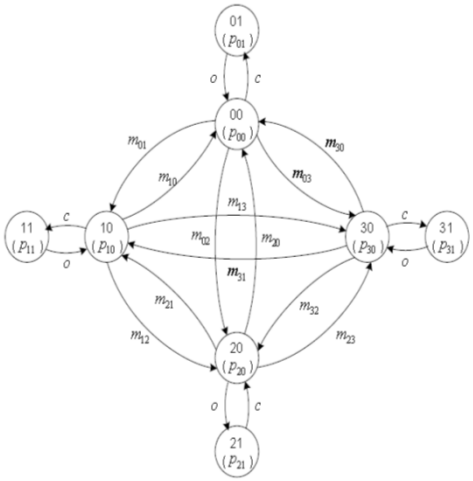




Predicate: Put
- Agent: Robot
- Patient: Apple (o1)
- Destination: Plate (o2)

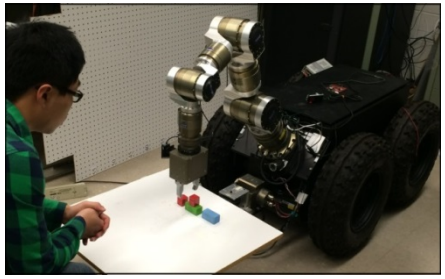
What controls actions of a robot?





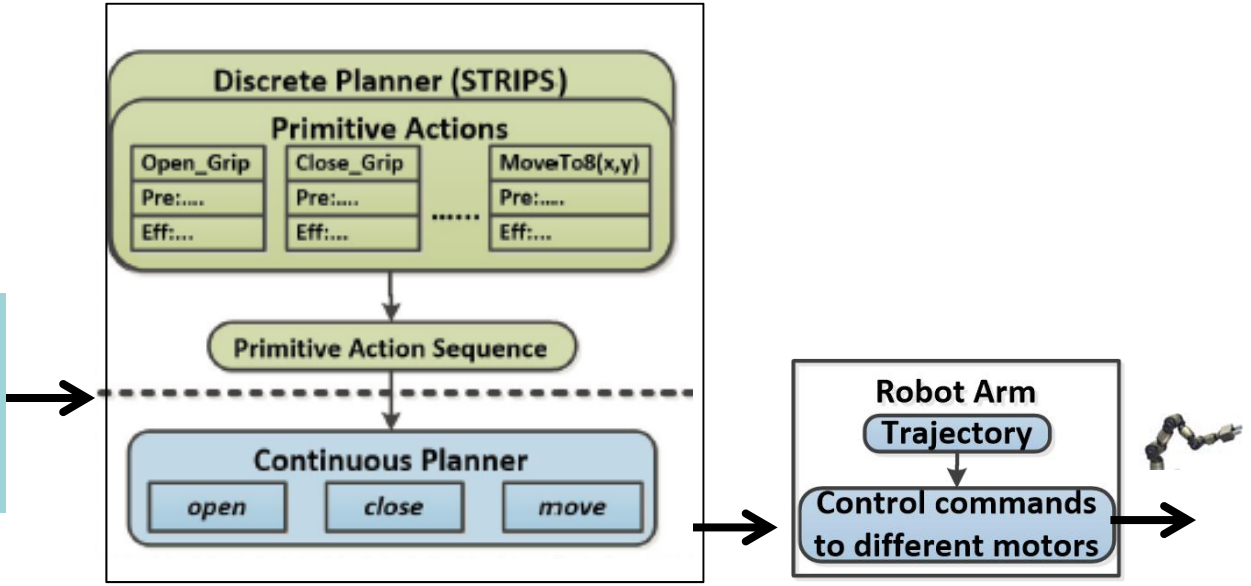
High-level actions specified by verbs will need to be translated to a sequence of primitive actions

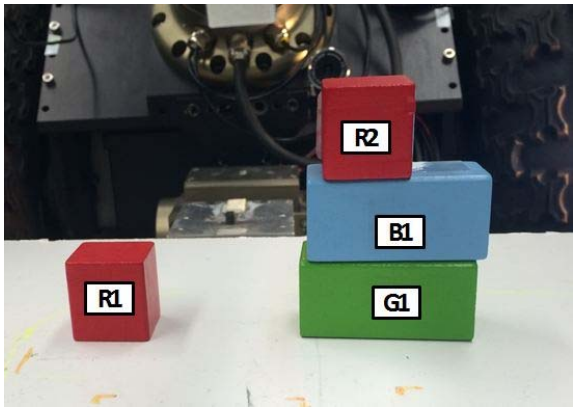
Representing Verbs with Goal State



What does "stack" mean?

Stack (A, B) : G_open
^ On(A, B) ^¬ On(A, Table)





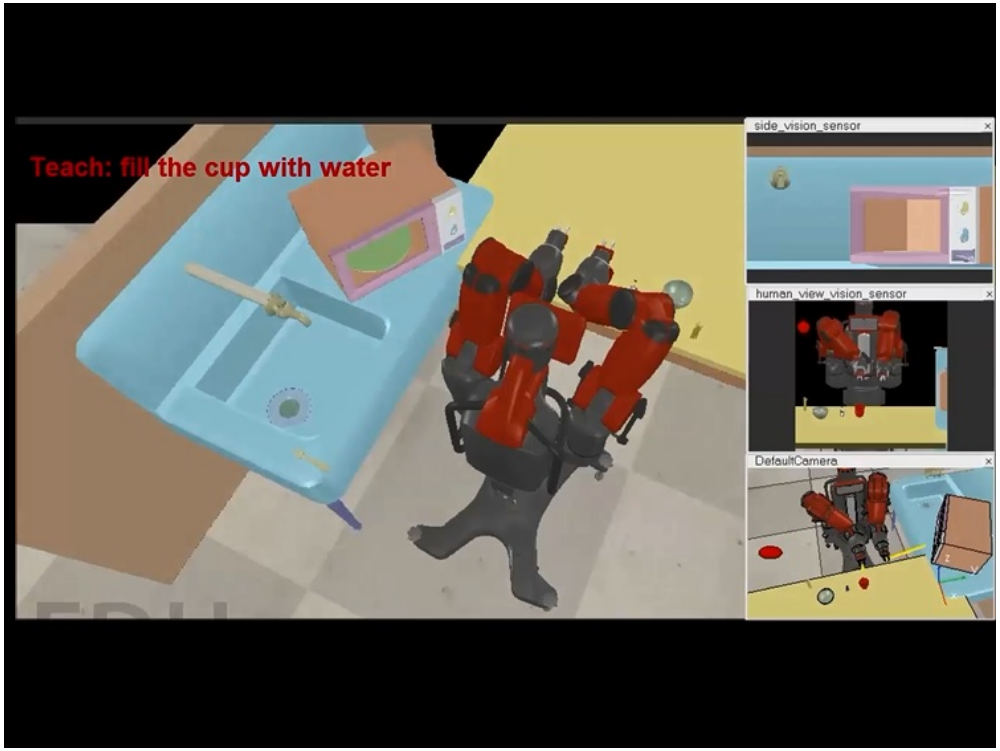
[stack(G1,R1)]

Automatically generated action sequence:

MoveTo(R2)-> Close_Grip->MoveTo(R2, TABLE)-> Open_Grip-> MoveTo(B1)-> Close_Grip-> MoveTo(B1,TABLE)-> Open_Grip-> MoveTo(G1)-> Close_Grip-> MoveTo(G1,R1)-> Open_Grip

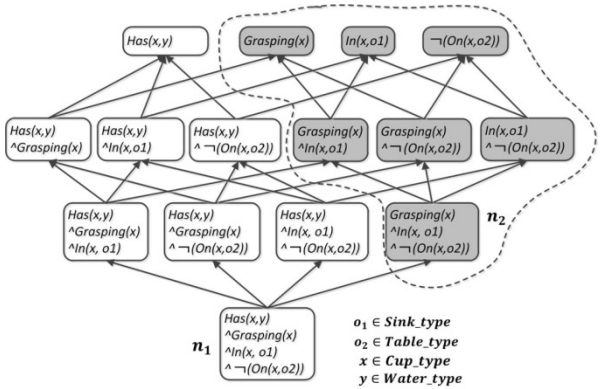
L. She, S. Yang, Y. Cheng, Y. Jia, J. Y. Chai, and N. Xi, *Back to the Blocks World: Learning New Actions through Situated Human-Robot Dialogue*, SIGDIAL 2014

Learning Verb Hypothesis Space



Acquire verb meanings through step –by-step language instructions with the user

Pick up the red cup
Turn to your left
Put the cup under the faucet
Turn on the faucet
Turn off the faucet



- The hypothesis space is incrementally updated (e.g., pruned and merged).
- Given a language command, the induced space together with the hypothesis selector can be applied by the agent to plan for lower-level actions.

L. She and J. Y. Chai. *Incremental Acquisition of Verb Hypothesis Space towards Physical World Interaction*, ACL 2016

A beginning of a long journey

Representation

Data and Knowledge

Learning and inference Methods

Collaboration