Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

Jesse Thomason

Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J. Mooney

The University of Texas at Austin

Grounded Linguistic Semantics

- Service robots are present in stores, factory floors, hospitals, and offices

- Need to understand language commands about the environment

Grounded Linguistic Semantics

- "Bring me the empty cup"

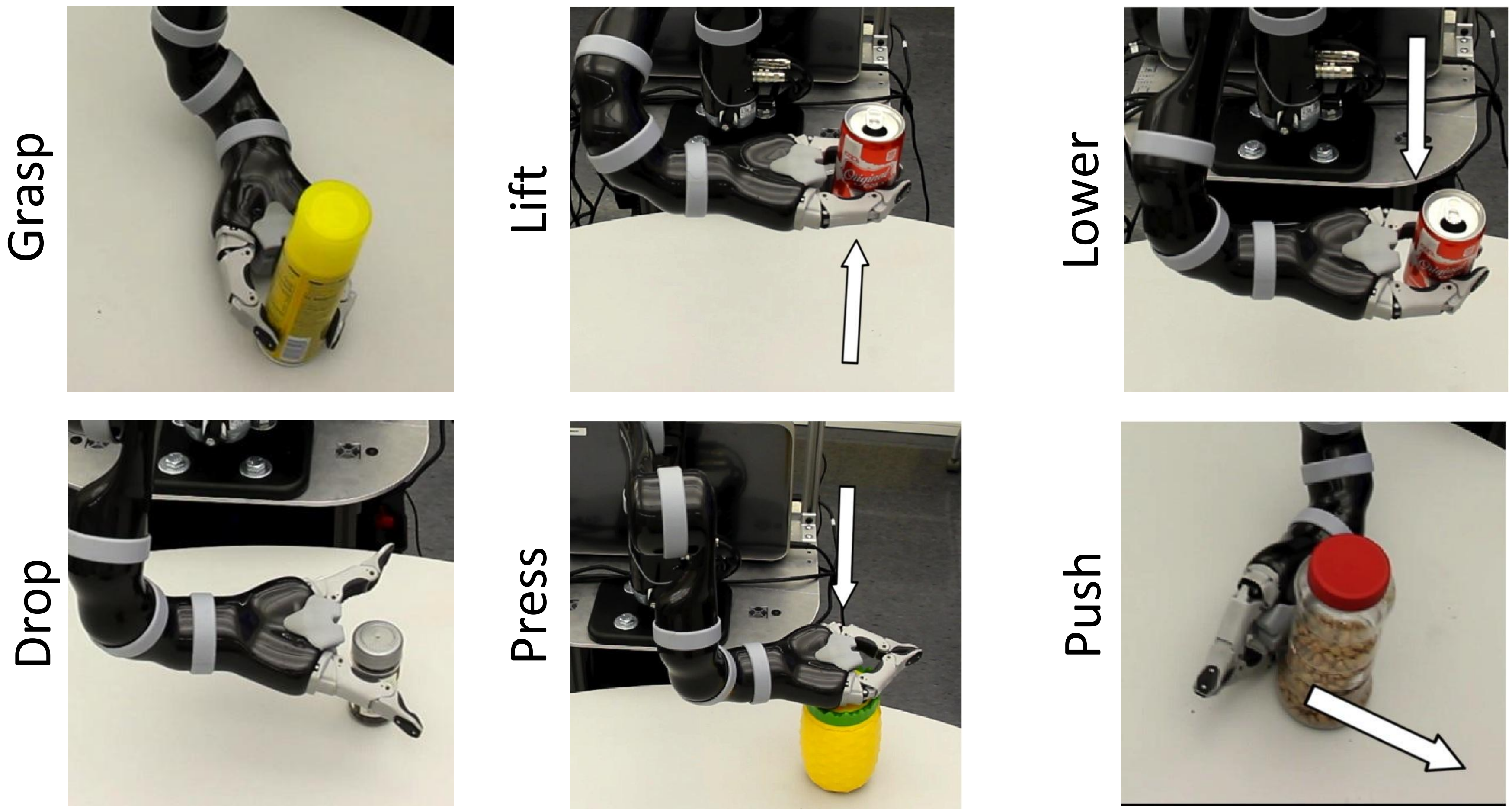
- Learn word meanings in terms of robot perception

Grounded Linguistic Semantics

- Traditionally done in vision space
- Predicates like “red” and “rectangle” can be learned through vision alone
- But looking isn’t all humans do
- “Empty”, “heavy”, “rattles”
- To understand some predicates, need to interact with objects beyond vision
- Equip a robot with both a camera and an arm

Multi-Modal
Grounded Linguistic Semantics

- Interact with objects beyond just looking



Multi-Modal Grounded Linguistic Semantics

- Represent objects with features from all *behaviors*
- Traditional and deep vision features from looking
- Audio, haptic, and proprioceptive features from manipulation behaviors
- Different types of features form sensory *modalities*

Multi-Modal Grounded Linguistic Semantics

- Every combination of *behavior* and *modality* forms an understanding *context*
- "Red" in the *look* + *color* context
- "Empty" in the *lift* + *haptic* context
- "Tall" in *look* + *shape*, *press* + *auditory* contexts
- Predicate classifiers composed of confidence-weighted votes from context classifiers

Learning Multi-Modal Grounded Linguistic Semantics

- Connect human language to features of sensory contexts
- Need labeled training data
  - This object is pink and short
- How do humans describe objects in question?
- Past work uses “I Spy” game (Parde 2015)

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

- Let the human and robot take turns describing objects
- Human descriptions give positive examples
- Robot descriptions followed up with dialog for positive and negative examples

"An empty metallic aluminum container"

"An empty metallic aluminum container"

Initially, robot has no training data and randomly guesses objects.

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

- System remembered positive and negative object examples for each predicate

empty container
metallic pink
aluminum yellow

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

- Train predicate classifiers from positive and negative object examples

empty: positive negative

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

- Predicate classifiers are a weighted vote of trained *context* classifiers giving decisions in [-1, 1] representing confidence

empty?

Behavior / Modality	color	…	audio	haptics
look	0.02		-	-
…	…	…	…	…
lift	-	…	-0.04	0.8
drop	-	…	0.4	0.02

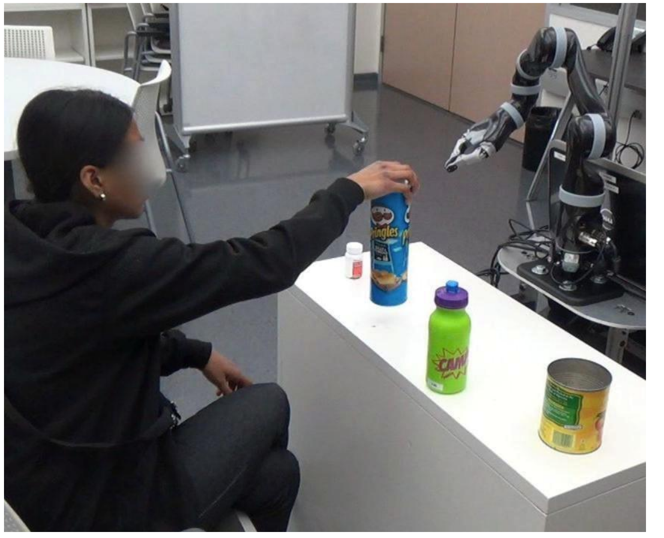
Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

- Use predicate classifiers confidences to decide how to describe a chosen object to the human

tub (+.8)
light (+.7)
tall (+.9)
pink (+.02)

short (-.8)
half-full (-.05)
empty (+.6)

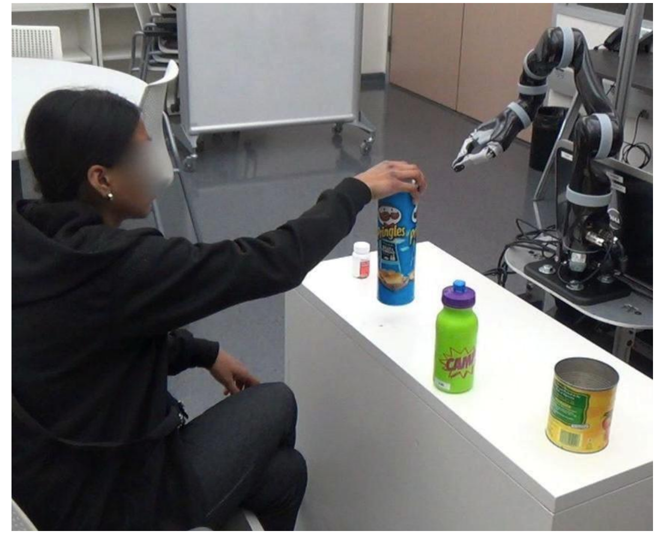
Robot Turn



"I am thinking of an object I would describe as light and tall and tub."

- Follow-up dialog gathers both positive and negative examples

Robot Turn

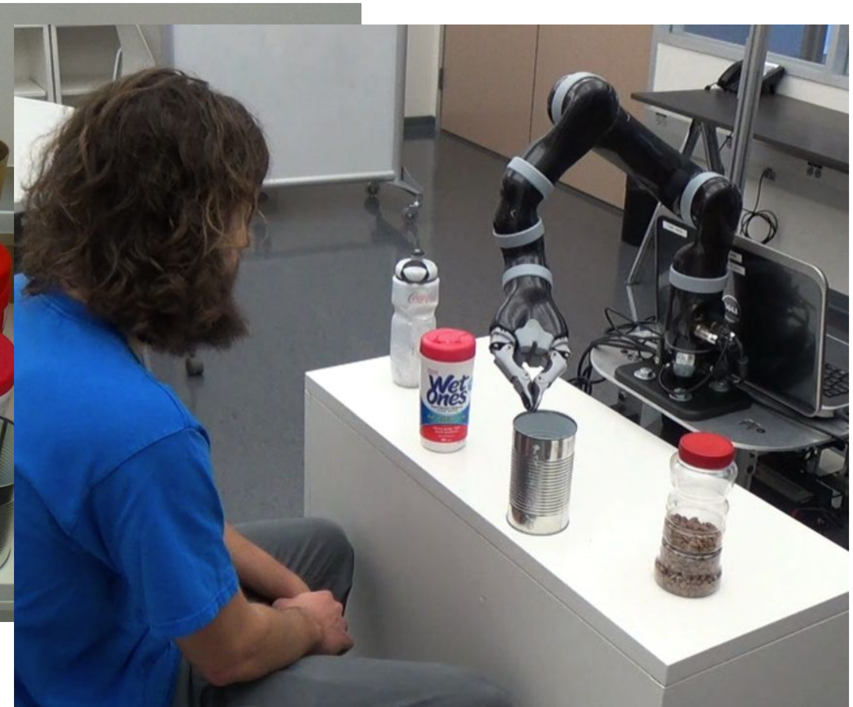


"Would you describe this object as light?"
"Would you describe this object as tall?"
"Would you describe this object as tub?"

"Would you describe this object as pink?"
"Would you describe this object as half-full?"

Playing "I Spy"

- Divided 32 objects into training folds of 8 each
- 10 participants played 4 games each with the robot; 4 objects per game



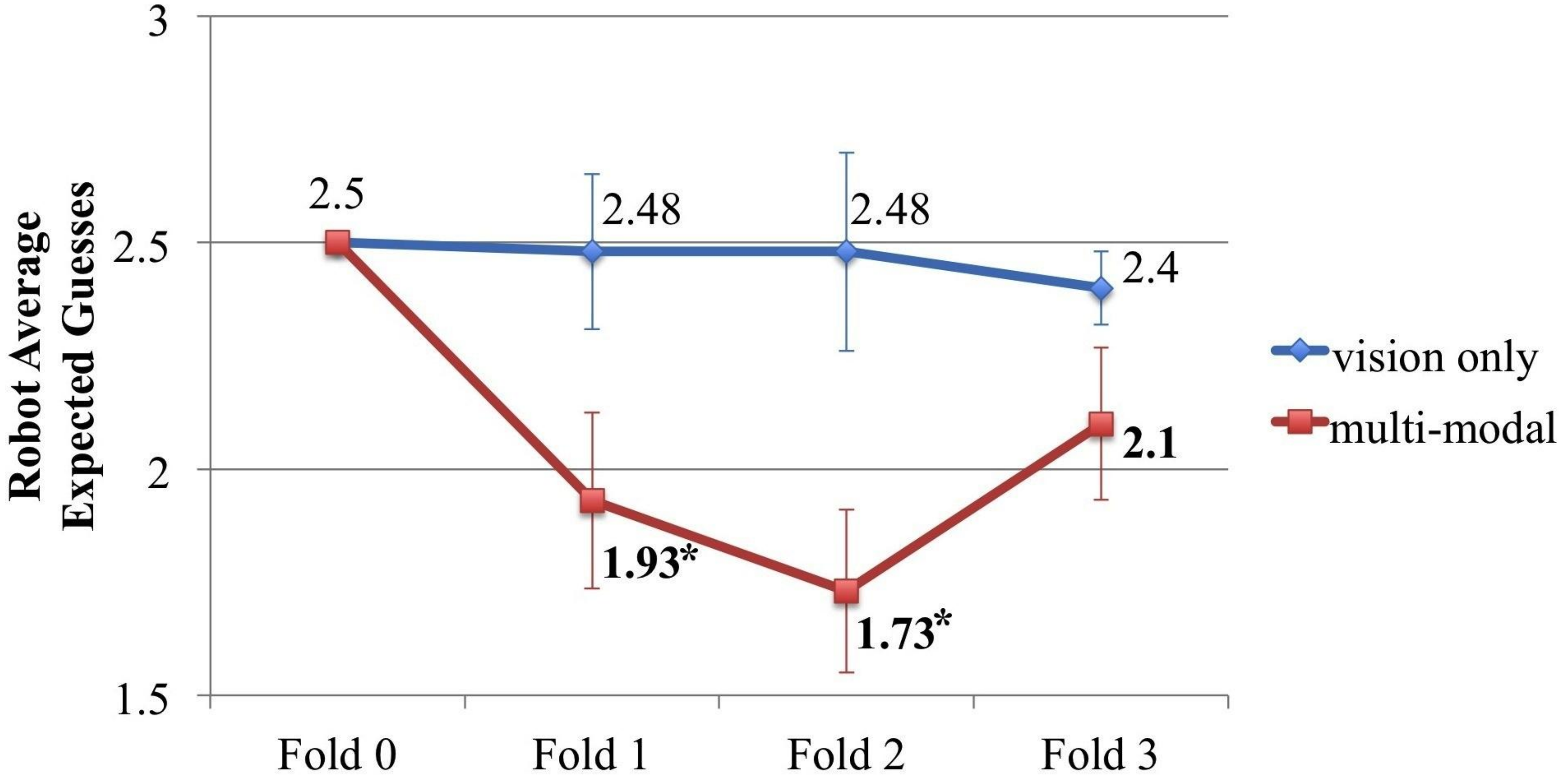
Playing "I Spy"

- Robot started with no vocabulary for first fold of 8 objects
- After each fold, learning phase allowed lexical acquisition and grounding
- Measured game performance on novel objects as more learning had taken place

Evaluating Multi-Modal Grounding

- Two learning algorithms compared
- **Vision only** baseline and **multi-modal** system
- During learning, **vision only** baseline only considered *look* behavior
- Users were unaware of multiple systems but interacted with both in 2 games each
  - All 8 objects seen by both systems per user
- Measured robot guesses for correct object

Results for Robot Guesses



Bold: Lower than fold 0 average. *****: Lower than vision only baseline

Results for Predicate Agreement

- Leave-one-object-out cross validation across predicate labels on objects (74 total learned)

Metric	System	
	vision only	multi-modal
precision	.250	.378+
recall	.179	.348*
F_1	.196	.354*

- *: significantly greater with $p < 0.05$
- +: trending greater with $p < 0.1$

Correlations to Physical Properties

- Pearson's *r* between predicate decision in [-1, 1] on object and height and weight
- **vision only** system learns no predicates with correlations $p < 0.05$ and $|r| > 0.5$
- **multi-modal** learns correlated predicates:
  - "tall" with height ($r = .521$)
  - "small" against weight ($r = -.665$)
  - "water" with weight ($r = .549$)

“A tall blue cylindrical container”

“A tall blue cylindrical container”

Conclusions

- We move beyond vision for grounding language predicates
- Auditory, haptic, and proprioceptive senses help understand words humans use to describe objects
- Some predicates assisted by multi-modal
  - "tall", "wide", "small"
- Some can be impossible without multi-modal
  - "half-full", "rattles", "empty"

Future Work

- Use one-class classification to remove need for negative examples
  - Move beyond "I Spy" to object retrieval alone
- Detect polysemy across modalities, as for the predicate "light" (color versus weight)
- Explore only as needed on novel objects
  - If predicate is "pink" with known relevant context *look* + *color*, only perform *look* behavior to decide

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

Jesse Thomason

Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J. Mooney

The University of Texas at Austin

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

https://youtu.be/jLHzRXPCi_w