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

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik,
Peter Stone, Raymond J. Mooney

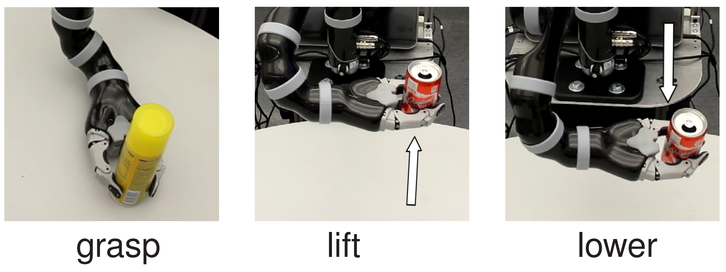
University of Texas at Austin

Multi-Modal Grounded Linguistic Semantics

Robots need to be able to connect language to their environment in order to discuss real world objects with humans. Mapping from referring expressions such as "the blue cup" to an object referent in the world is an example of the *symbol grounding problem* (Harnad 1990). Symbol grounding involves connecting internal representations of information in a machine to real world data from its sensory perception. *Grounded language learning* bridges these symbols with natural language.
We refer to adjectives and nouns that describe properties of objects as language *predicates*. Most work has focused on grounding predicates through visual information. However, other sensory modalities such as haptic and auditory are also useful in allowing robots to discriminate between object categories (Sinapov 2014b).

Multi-Modal Sensory Perception

We ground language predicates by considering visual, haptic, auditory, and proprioceptive senses. The robot used in this study was a Kinova MICO arm mounted on top of a custom-built mobile base whose perception included joint effort sensors in each of the robot arm's motors, a microphone mounted on the mobile base, and an Xtion ASUS Pro RGBD camera.



Behavior	Modality		
	color	fpfh	vgg
look	64	308	4096
	audio	haptics	proprioception
grasp	100	60	20
drop, hold, lift, lower, press, push	100	60	

Left: The behaviors the robot used to explore the objects. The arrows indicate the direction of motion of the end-effector for each behavior. In addition, the *hold* behavior (not shown) was performed after the *lift* behavior by simply holding the object in place for half a second. **Right:** The number of features extracted from each *context*, or combination of robot behavior and perceptual modality.

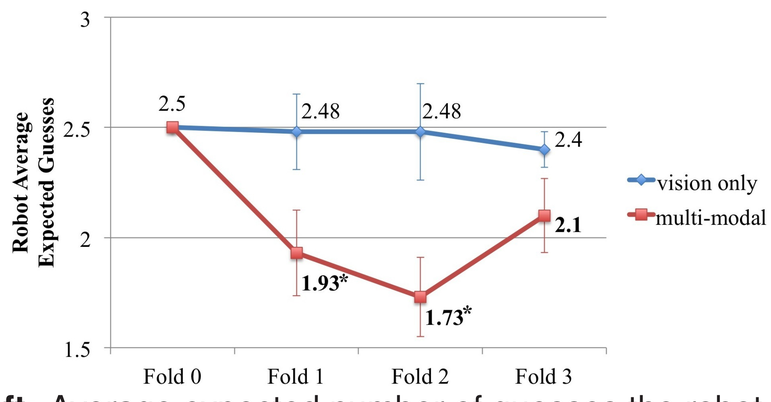
Playing "I Spy"

A home or office robot can explore objects in an unsupervised way to gather perceptual data, but needs human supervision to connect this data to language. Learning grounded semantics through human-robot dialog allows a system to acquire the relevant knowledge without the need for laborious labeling of numerous objects for every potential lexical descriptor. A few groups have explored learning from interactive linguistic games such as "I Spy" and "20 Questions" (Parde 2015, Vogel 2010); however, these studies only employed vision. We use a variation on the children's game "I Spy" as a learning framework for gathering human language labels for objects to learn multi-modal grounded lexical semantics.

Left: Objects used in the "I Spy" game divided into four folds, from fold 0 on the left to fold 3 on the right. **Center**: the robot guesses an object described by a human participant as "silver, round, and empty." **Right**: a human participant guesses an object described by the robot as "light," "tall," and "tub."

Experiment

To determine whether multi-modal perception helps a robot learn grounded language, we had two different systems play "I Spy" with 42 human participants and measured performance as more training data from previous games became available. The baseline **vision only** system used only the *look* behavior when grounding language predicates, while our **multi-modal** system used the full suite of behaviors and associated haptic, proprioceptive, and auditory modalities.



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

Left: Average expected number of guesses the robot made on each human turn. **Bold**: lower than fold 0 with $p < 0.05$. *: lower than the competing system on participant-by-participant basis with $p < 0.05$. **Right:** Average performance of predicate classifiers in leave-one-object-out cross validation. *: greater than competing system with $p < 0.05$. **+**: $p < 0.1$.

Predicate	$f_1^{mm} - f_1^{vo}$	High Confidence Positive	High Confidence Negative
		multi-modal system	
can	0.857		
tall	0.516		
half-full	.462		
yellow	.312		
		vision only system	
pink	-.3		

Sample of predicates for which performance between the systems was substantially different. The highest- and lowest-confidence objects for each predicate are shown. The top rows ($f_1^{mm} - f_1^{vo} > 0$) are decisions from the **multi-modal** system, the bottom row from the **vision only** system.

We also calculated the Pearson's correlation r between predicate decisions on each object and objects' weights, heights, and widths. The **vision only** system led to no predicates substantially and significantly correlated against these physical object features. The **multi-modal** "tall" predicate correlates with objects that are higher ($r = .521$), "small" ($r = -.665$) correlates with objects that are lighter, and "water" ($r = .814$) correlates with objects that are heavier. The latter is likely from objects described as "water bottle", which, in our dataset, are mostly filled either half-way or totally and thus heavier. There is also a spurious correlation between "blue" and weight ($r = .549$).

Conclusion

We expand past work on grounding natural language in robot sensory perception by going beyond vision and exploring haptic, auditory, and proprioceptive robot senses. We compare a vision only grounding system to one that uses these additional senses by employing an embodied robot playing "I Spy" with many human users. To our knowledge, ours is the first robotic system to perform natural language grounding using multi-modal sensory perception through natural interaction with human users.

jesse@cs.utexas.edu https://www.cs.utexas.edu/~ml/