Guiding Interaction Behaviors for Multi-modal Grounded Language Learning

Jesse Thomason, Jivko Sinapov & Raymond J. Mooney

Presented by Siliang Lu

Multi-modal grounded language learning

Multiple modalities

Language predicates → Physical properties of objects in the world

Interaction behaviors

Visual predicate (i.e. “red”)

Non-visual predicate (i.e. “empty”)

Modalities: Audio, Haptics, visual colors and shapes

Behaviors: look, drop, grasp, hold, lift, lower, press, push

Classification

Weighting scheme:

- Only validation confidence ($\kappa, Cohen's\ kappa\ agreements$) with human labels using leave-one-out cross-validation
- Confidence and behavior annotations
- Confidence and modality annotations
- Confidence and word similarity

Consideration of only validation confidence

Method:

- SVM using the feature space for each sensorimotor context (a combination of a behavior and sensory modality)

Behaviors	**Modalities**
look	color, fpfh
drop, grasp, hold, lift lower, press, push	audio, haptics

Sensorimotor context

Consideration of only validation confidence

Decision $d(p, o) \in [-1, 1]$ for predicate p and object o is defined as:

$$d(p, o) = \sum_{c \in C} \kappa_{p,c} G_{p,c}(o) \geq 0$$

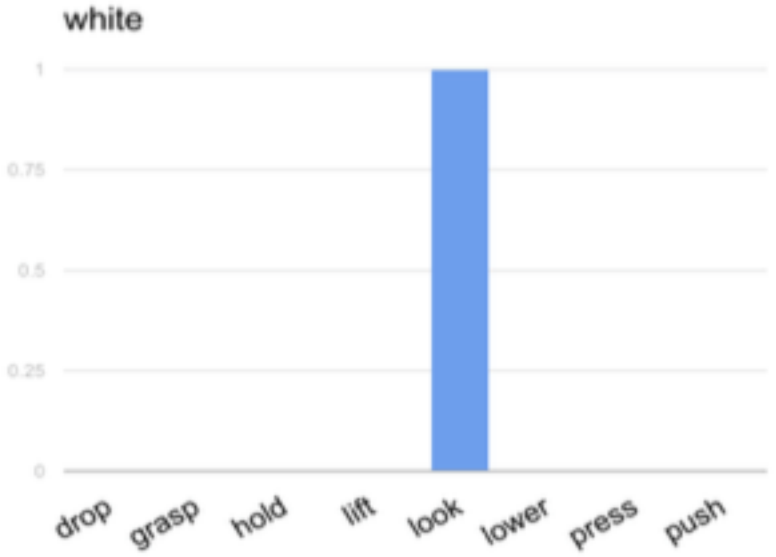
κ: Cohen's kappa agreement where the higher value indicates larger influence.
i.e. examples for "red" in "look/color" space is weighed higher than in "drop/audio" space

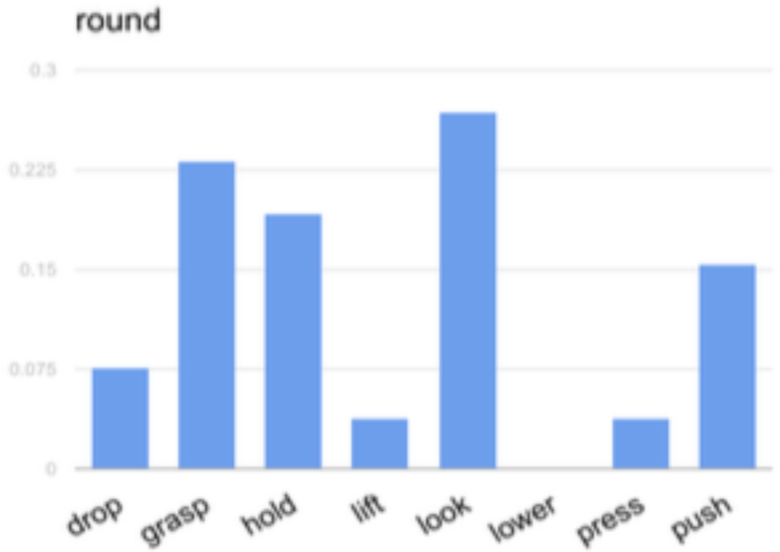
$G_{p,c}$: a supervised grounding classifier (i.e. SVM with linear kernel)
trained on labeled object, which returns {-1,1}

Confidence and behavior annotations

Interaction behavior annotations:

- Manual labelling by asking which exploratory behaviors annotators would engage in.
- Among 14 annotators, 8 of them with higher average kappa agreement than 0.4 were chosen.
- Induction of a distribution over behaviors b ∈ B
- A^{B}_{p,C_b}: Proportion of annotators who marked behavior *b* relevant for understanding predicate *p*





Confidence and behavior annotations

$$d(p, o) = \sum_{c \in C} A^{B}_{p,c_b} \kappa_{p,c} G_{p,c}(o) \geq 0$$

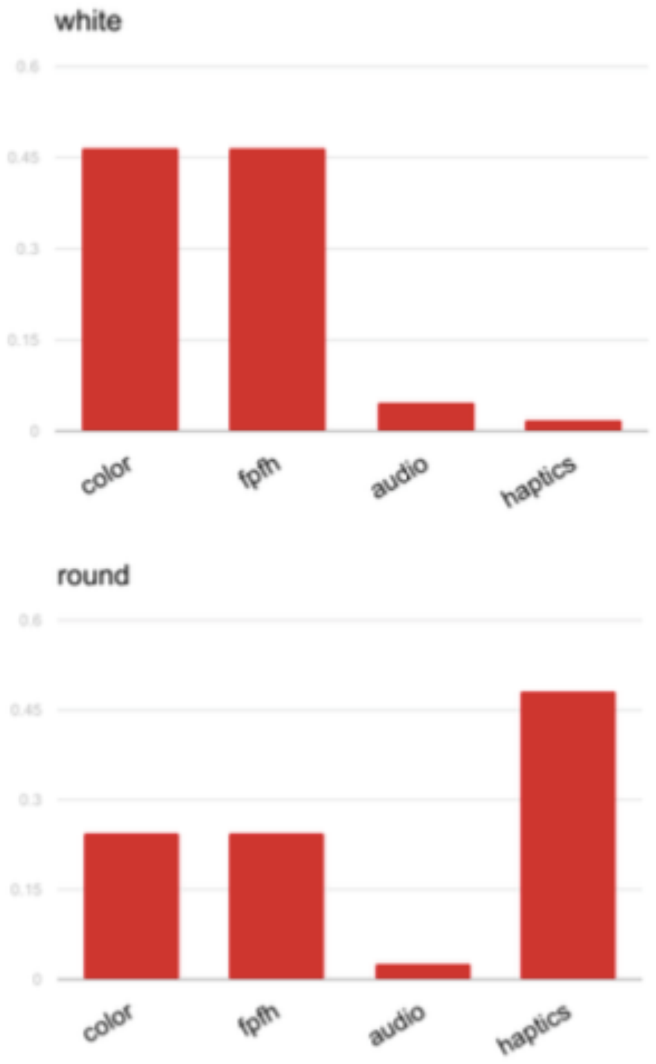
A^{B}_{p,c_b}: Proportion of annotators who marked behavior *b* relevant for understanding predicate *p*

Confidence and multi-modality annotations

$$d(p,o) = \sum_{c \in C} A^{M}_{p,C_m} \kappa_{p,c} G_{p,c}(o) \geq 0$$

A^{M}_{p,C_b}*: Proportion of the modality exclusivity norm marking behavior *b* relevant for understanding predicate *p*, which gathered from past work

Modalities: auditory, haptic, visual color and visual shapes



* When A^{M}_{p,C_m} is not in the past work, a uniform 1/|M| is used.

Sharing confidence between related predicates

- Calculating cosine distance in word embedding space by using **Word2Vec**

For every pair of predicates $p,q \in P$ with word embedding vectors v_p, v_q, the similarity can be calculated as:

$$w(p,q) = \frac{1}{2}\left(1 + \cos\left(v_p, v_q\right)\right) \in [0,1]$$

Sharing confidence between related predicates

$$d(p,o) = \sum_{c \in C} \left(|P|^{-1} \sum_{q \in P} w(p,q)\, \kappa_{q,c} \right) G_{p,c}(o) \geq 0$$

i.e. if kappa of “thin, grasp/haptic” is high for the predicate “narrow”, we should trust grasp/haptic sensorimotor context

Results

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

F1-measurement$\frac{2}{F1} = \frac{1}{Precision} + \frac{1}{Recall}$

Results

- Adding behavior annotations or modality annotations improves performance over using kappa alone
- Sharing kappa information improves recall at the cost of precision
  - Trade-off due to real world “noise” in specific domains.
  - *i.e.* “water” correlated with object weights

	p	r	f1
mc	.282	.355	.311
κ	.406	.460	.422
B+κ	**.489**	**.489**	**.465**
M+κ	.414	.466	.430
W+κ	.373	.474	.412

Future work

- Apply behavior annotations in an embodied dialog agent
- Explore other methods of sharing information between predicates such as using a maximally similar neighbor word
  - i.e. the best neighbor of “narrow” is “thin”

Thanks!