In Proceedings of the Workshop on Language Grounding for Robotics at ACL 2017 (RoboNLP-17), Vancouver, Canada, August 2017.

Guiding Interaction Behaviors for Multi-modal Grounded Language Learning

Jesse Thomason, Jivko Sinapov, and **Raymond J. Mooney**
Department of Computer Science, University of Texas at Austin
Austin, TX 78712, USA
{jesse, jsinapov, mooney}@cs.utexas.edu

Abstract

Multi-modal grounded language learning connects language predicates to physical properties of objects in the world. Sensing with multiple modalities, such as audio, haptics, and visual colors and shapes while performing interaction behaviors like lifting, dropping, and looking on objects enables a robot to ground non-visual predicates like "empty" as well as visual predicates like "red". Previous work has established that grounding in multi-modal space improves performance on object retrieval from human descriptions. In this work, we gather behavior annotations from humans and demonstrate that these improve language grounding performance by allowing a system to focus on relevant behaviors for words like "white" or "half-full" that can be understood by looking or lifting, respectively. We also explore adding modality annotations (whether to focus on audio or haptics when performing a behavior), which improves performance, and sharing information between linguistically related predicates (if "green" is a color, "white" is a color), which improves grounding recall but at the cost of precision.

1 Introduction

Connecting human language predicates like "red" and "heavy" to machine perception is part of the *symbol grounding problem* (Harnad, 1990), approached in machine learning as *grounded language learning*. For many years, grounded language learning has been performed primarily in visual space (Roy and Pentland, 2002; Liu et al., 2014; Malinowski and Fritz, 2014; Mohan et al., 2013; Sun et al., 2013; Dindo and Zambuto, 2010; Vogel et al., 2010). Recently, researchers have explored grounding in audio (Kiela and Clark, 2015), haptic (Alomari et al., 2017), and multi-modal (Thomason et al., 2016) spaces. Multi-modal grounding allows a system to connect language predicates like "rattles", "empty", and "red" to their audio, haptic, and color signatures, respectively.

Past work has used human-robot interaction to gather language predicate labels for objects in the world (Parde et al., 2015; Thomason et al., 2016). Using only human-robot interaction to gather labels, a system needs to learn effectively from only a few examples. Gathering audio and haptic perceptual information requires doing more than looking at each object. In past work, multiple interaction behaviors are used to explore objects and add this audio and haptic information (Sinapov et al., 2014).

In this work, we gather annotations on what exploratory behaviors humans would perform to determine whether language predicates apply to a novel object. A robot could gather such information by asking human users which action would best allow it to test a particular property, e.g. "To tell whether something is 'heavy' should I look at it or pick it up?" Figure 1 shows some of the behaviors used by our robot in previous work to perceive objects and their properties. In this paper, we show that providing a language grounding system with behavior annotation information improves classification performance on whether predicates apply to objects, despite having sparse predicate-object labels.

We additionally explore adding modality annotations (e.g. is a predicate more auditory or more haptic), drawing on previous work in psychology that gathered modality norms for many words (Lynott and Connell, 2009). Finally, we explore using

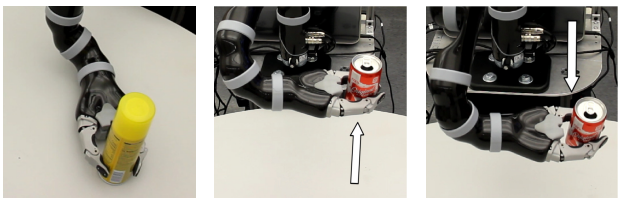

Figure 1: Behaviors the robot used to explore objects. In addition, the *hold* behavior (not shown) was performed after the *lift* behavior by holding the object in place for half a second. The *look* behavior (not shown) was also performed for all objects.

word embeddings to help with infrequently seen predicates by sharing information with more common ones (e.g. if "thin" is common and "narrow" is rare, we can exploit the fact that they are linguistically related to help understand the latter).

2 Dataset and Methodology

Previous work provides sparse annotations of 32 household objects (Figure 2) with language predicates derived during an interactive "I Spy" game with human users (Thomason et al., 2016). Each predicate $p \in P$ from that work is associated with objects as applying or not applying, based on dialog with human users. For example, predicate "red" applies to several objects and not to others, but for many objects its label is not explicitly known. Objects are represented by features gathered during several interaction behaviors (Figure 1) as detailed in past work (Sinapov et al., 2016). In this work, we focus on improving the language grounding performance of multimodal classifiers that predict whether each predicate $p \in P$ applies to each object $o \in O$.

In previous work, decisions about a predicate and an object are made for each sensorimotor context (a combination of a behavior and sensory modality) with an SVM using the feature space for that context (Thomason et al., 2016). A summary of sensorimotor contexts is given in Table 1.

Figure 2: Objects explored via interaction behaviors and for which we have sparse predicate annotations.

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

Table 1: The *contexts* (combinations of robot behavior and perceptual modality) we use for multimodal language grounding. The *color* modality is color histograms, *fpfh* is fast-point feature histograms, *audio* is fast Fourier transform frequency bins, and *haptics* is averages over robot arm joint forces (detailed in (Sinapov et al., 2016)).

For example, a classifier is trained from the positive and negative object examples for "red" in *look/color* space as well as in the less relevant *drop/audio* space. These decisions are then averaged together, each weighted by its Cohen's-κ agreement with human labels using leave-one-out cross validation on the training data. In this way, the *look/color* space for "red" is expected to have high κ and a large influence on the decision, while *drop/audio* would have low κ and not influence the decision much.

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

$$d(p, o) = \sum_{c \in C} \kappa_{p,c} G_{p,c}(o), \qquad (1)$$

for $G_{p,c}$ a supervised grounding classifier trained on labeled objects for predicate p in the feature space of sensorimotor context c that returns in $\{-1, 1\}$ with $\kappa_{p,c}$ its agreement with human labels. If $d(p, o) \leq 0$, we say p does not apply to o, else that it does. We use SVMs with linear kernels as grounding classifiers.

We extend the weighting scheme between sensorimotor SVMs to include behavior information. For each predicate derived from the "I Spy" game

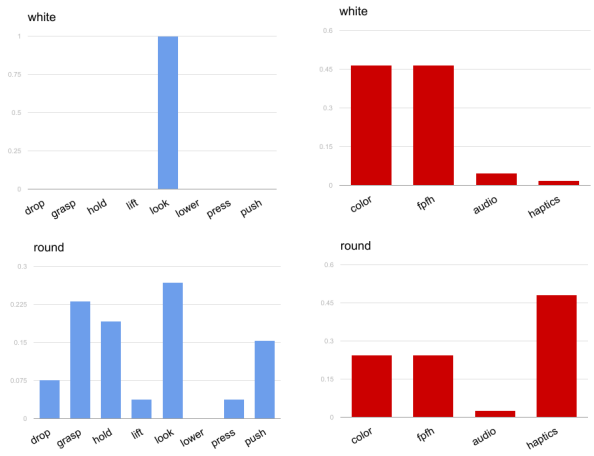

Figure 3: The distribution over annotator-chosen behaviors (left) gathered in this work, as well as the distribution over modality norms (right) derived from previous work (Lynott and Connell, 2009), for the predicates "white" and "round". The *fpfh* modality is fast-point feature histograms.

in previous work, we gather relevant behaviors from human annotators. Annotators were asked to mark which among the 8 exploratory behaviors (Table 1) they would engage in to determine whether a given predicate applied to a novel object. Annotators could mark as many behaviors as they wanted for each predicate, but were required to choose at least one.

We gathered annotations from 14 people, then discarded the annotations from those whose average κ agreement with all other annotators was less than 0.4 (the poor-fair agreement range). This left us with 8 annotators whose average $\kappa = .475$ (moderate agreement). We release the full set of gathered annotations on 81 perceptual predicates across 8 behaviors as a corpus for community use.[1]

Then, for each $p \in P$, we induce a distribution over behaviors $b \in B$ based on the ratio of annotators that marked that behavior relevant, such that $\sum_{b \in B} A^B_{p,b} = 1$, with $A^B_{p,b}$ equal to the proportion of annotators who marked behavior b relevant for understanding predicate p. Some predicates, like "white", have single behavior distributions. For other predicates, like "metal", annotators chose more complex combinations of behaviors. Figure 3 (Left) gives some examples of behavior distributions from our annotations.

[1] http://jessethomason.com/publication_supplements/robonlp_thomason_mooney_behavior_annotations.csv

The decision $d_B(p, o)$ considering behavior annotations is calculated as

$$d_B(p, o) = \sum_{c \in C} A^B_{p,c_b} \kappa_{p,c} G_{p,c}(o), \qquad (2)$$

where c_b is the behavior for sensorimotor context c.

We also experiment with adding modality annotations (Table 1). In particular, we derive a modality distribution for each $p \in P$ such that $\sum_{m \in M} A^M_{p,m} = 1$ from modality exclusivity norms gathered by past work for auditory, gustatory, haptic, olfactory, and visual modalities (Lynott and Connell, 2009). We ignore gustatory and olfactory modalities, which have no counterpart in our sensorimotor contexts, and create $A^M_{p,m}$ scores from the auditory, haptic, and visual modality norm means. The visual modality norm is split evenly between relevance scores $A^M_{p,\text{color}}$ and $A^M_{p,\text{fpfh}}$, our visual color and shape modalities.

The decision $d_M(p, o)$ considering modality annotations is calculated as

$$d_M(p, o) = \sum_{c \in C} A^M_{p,c_b} \kappa_{p,c} G_{p,c}(o) \qquad (3)$$

When the predicate p does not appear in the norming dataset from past work[2], a uniform $A^M_{p,m} = 1/|M|$ is used. Figure 3 (Right) gives some examples of modality distributions from these norms.

The data sparsity inherent in language grounding from limited human interaction means some predicates have just a handful of positive and negative examples, while more common predicates may have many. If we have few examples for "narrow" but many for "thin," we can share some information between them. For example, if $\kappa_{\text{thin,grasp/haptic}}$ is high, we should trust the *grasp/haptic* sensorimotor context for "narrow" more than "narrow"s κ estimates alone suggest.

We explore sharing κ information between related predicates by calculating their cosine distance in word embedding space by using Word2Vec (Mikolov et al., 2013) vectors derived from Google News.[3] For every pair of predicates $p, q \in P$ with word embedding vectors v_p, v_q we calculate similarity as

$$w(p, q) = \frac{1}{2}(1 + cos(v_p, v_q)), \qquad (4)$$

[2] About half the predicates have norming information.

[3] https://github.com/mmihaltz/word2vec-GoogleNews-vectors

	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

Table 2: Precision (**p**), recall (**r**), and $f1$ (**f1**) of predicate classifiers across weighting schemes. **mc** gives majority class baseline. Weighting schemes consider only validation confidence (κ, as in previous work), confidence and behavior annotations (**B**+κ), confidence and modality annotations (**M**+κ), and confidence and word similarity (**W**+κ). Note that we show the average per-predicate f-measure, not the f-measure of the average per-predicate precision and recall.

which falls in $[0, 1]$, and subsequently take a weighted average of κ values using these similarities as weights to get decisions $d_W(p, o)$ as

$$d_W(p,o) = \sum_{c \in C} \left(|P|^{-1} \sum_{q \in P} \kappa_{q,c} w(p,q) \right) G_{p,c}(o) \quad (5)$$

3 Experimental Evaluation

We calculated precision, recall, and f-measure between human labels and predicate decisions when weighting constituent sensorimotor context classifiers by the schemes described above: kappa confidence only (Eq 1, κ), adding behavior annotations (Eq 2, **B**+κ), adding modality annotations (Eq 3, **M**+κ), and sharing kappas across predicates using word similarity (Eq 5, **W**+κ).

We calculated these metrics for each predicate[4] and averaged scores across all predicates. We use leave-one-object-out cross validation to obtain performance statistics for each weighting scheme.

Table 2 gives the results for predicates that have at least 3 positive and 3 negative training object examples.[5]

We observe that adding behavior annotations or modality annotations improves performance over using kappa confidence alone, as was done in past work. Sharing kappa confidences across similar predicates based on their embedding cosine similarity improves recall at the cost of precision.

Adding behavior annotations helps more than adding modality norms, but we gathered behavior annotations for all predicates, while modality annotations were only available for a subset (about half). Adding behavior annotations helped the f-measure of predicates like "pink", "green", and "half-full", while adding modality annotations helped with predicates like "round", "white", and "empty".

Sharing confidences through word similarity helped with some predicates, like "round", at the expense of domain-specific meanings of predicates like "water". In the "I Spy" paradigm from which these data were gathered, the authors noted that "water" correlated with object weight because all of their water bottle objects were partially or completely full (Thomason et al., 2016). Thus, in that domain, "water" is synonymous with "heavy". In a less restricted domain, word similarity may add less real world "noise" to the problem.

4 Conclusions and Future Work

In this work, we have demonstrated that behavior annotations can improve language grounding for a platform with multiple interaction behaviors and modalities. In the future, we would like to apply this intuition in an embodied dialog agent. If a person asks a service robot to "Get the white cup.", the robot should be able to ask "What should I do to tell if something is 'white'?", a behavior annotation prompt. A human-robot POMDP dialog policy could be learned, as in previous work (Padmakumar et al., 2017), to know when this kind of follow-up question is warranted.

Additionally, we will explore other methods of sharing information between predicates from lexical information. For example, choosing a maximally similar neighboring word, rather than doing a weighted average across all known words, may yield better results (e.g. the best neighbor of "narrow" is "thin", so don't bother considering things like "green" at all).

Acknowledgments

We thank our anonymous reviewers for their time and insights. This work is supported by a National Science Foundation Graduate Research

[4] Decisions were made for each testing object and marked correct or incorrect against human labels that object, if available for the predicate.

[5] The trends are similar when considering all predicates, but the scores and differences in performance are lower due to many predicates having only a single positive or negative example.

Fellowship to the first author, an NSF EAGER grant (IIS-1548567), and an NSF NRI grant (IIS-1637736). A portion of this work has taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (CNS-1330072, CNS-1305287, IIS-1637736, IIS-1651089), ONR (21C184-01), AFOSR (FA9550-14-1-0087), Raytheon, Toyota, AT&T, and Lockheed Martin.

References

Muhannad Alomari, Paul Duckworth, David C. Hogg, and Anthony G. Cohn. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. pages 4349–4356.

Haris Dindo and Daniele Zambuto. 2010. A probabilistic approach to learning a visually grounded language model through human-robot interaction. In *International Conference on Intelligent Robots and Systems*. IEEE, Taipei, Taiwan, pages 760–796.

S. Harnad. 1990. The symbol grounding problem. *Physica D* 42:335–346.

Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Emperical Methods in Natural Language Processing*. Lisbon, Portugal, pages 2461–2470.

Changson Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. 2014. Probabilistic labeling for efficient referential grounding based on collaborative discourse. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland, USA, pages 13–18.

Dermot Lynott and Louise Connell. 2009. Modality exclusivity norms for 423 object properties. *Behavior Research Methods* 41(2):558–564.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*. Montréal, Canada, pages 13–18.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Lake Tahoe, Nevada, pages 3111–3119.

Shiwali Mohan, Aaron H. Mininger, and John E. Laird. 2013. Towards an indexical model of situated language comprehension for real-world cognitive agents. In *Proceedings of the 2nd Annual Conference on Advances in Cognitive Systems*. Baltimore, Maryland, USA.

Aishwarya Padmakumar, Jesse Thomason, and Raymond J. Mooney. 2017. Integrated learning of dialog strategies and semantic parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain, pages 547–557.

Natalie Parde, Adam Hair, Michalis Papakostas, Konstantinos Tsiakas, Maria Dagioglou, Vangelis Karkaletsis, and Rodney D. Nielsen. 2015. Grounding the meaning of words through vision and interactive gameplay. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. Buenos Aires, Argentina, pages 1895–1901.

Deb Roy and Alex Pentland. 2002. Learning words from sights and sounds: a computational model. *Cognitive Science* 26(1):113–146.

Jivko Sinapov, Priyanka Khante, Maxwell Svetlik, and Peter Stone. 2016. Learning to order objects using haptic and proprioceptive exploratory behaviors. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. 2014. Learning relational object categories using behavioral exploration and multimodal perception. In *IEEE International Conference on Robotics and Automation*.

Yuyin Sun, Liefeng Bo, and Dieter Fox. 2013. Attribute based object identification. In *International Conference on Robotics and Automation*. IEEE, Karlsruhe, Germany, pages 2096–2103.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond Mooney. 2016. Learning multi-modal grounded linguistic semantics by playing "I spy". In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. pages 3477–3483.

Adam Vogel, Karthik Raghunathan, and Dan Jurafsky. 2010. Eye spy: Improving vision through dialog. In *Association for the Advancement of Artificial Intelligence*. pages 175–176.