Guiding Exploratory Behaviors for Multi-Modal Grounding of Linguistic Descriptions

Jesse Thomason[1] **Jivko Sinapov**[2] **Raymond J. Mooney**[1] **Peter Stone**[1]

[1]Computer Science Dept.
University of Texas at Austin
jesse, mooney, pstone@cs.utexas.edu

[2]Computer Science Dept.
Tufts University
jsinapov@cs.tufts.edu

Abstract

A major goal of grounded language learning research is to enable robots to connect language predicates to a robot's physical interactive perception of the world. Coupling object exploratory behaviors such as grasping, lifting, and looking with multiple sensory modalities (e.g., audio, haptics, and vision) enables a robot to ground non-visual words like "heavy" as well as visual words like "red". A major limitation of existing approaches to multi-modal language grounding is that a robot has to exhaustively explore training objects with a variety of actions when learning a new such language predicate. This paper proposes a method for guiding a robot's behavioral exploration policy when learning a novel predicate based on known grounded predicates and the novel predicate's linguistic relationship to them. We demonstrate our approach on two datasets in which a robot explored large sets of objects and was tasked with learning to recognize whether novel words applied to those objects.

Introduction

The symbol grounding problem is that of connecting natural language phrases to objects in the real world (Harnad 1990). To perform this task, a robot must represent words like "green" and "heavy" not in terms of other words (like dictionary definitions), but in terms of its own sensory perception. This is the task of *grounded language learning*. Past work has demonstrated that using sensory data beyond vision to ground language predicates improves robotic performance over using vision alone (Thomason et al. 2016). Non-visual exploratory behaviors such as pushing, grasping, and lifting the object can be costly in terms of time (e.g. localizing an object with a camera in order to press down on it) and operator intervention (e.g. pushing a ball off a table, requiring an operator to retrieve it).

In many settings, a robot needs to perform object exploration for a specific grounding task. For example, if someone asks a household robot to "get the heavy mug from the kitchen," the robot may need to explore some novel objects in the kitchen to determine which one satisfies "heavy" and "mug." If a dataset of unexplored objects labeled as "heavy" (or not) and "mug" (or not) is available, the robot should be able to explore those objects quickly to learn the concepts before traveling to the kitchen to identify the referent. Exhaustively performing all actions when exploring new objects to learn a novel predicate will, in expectation, yield the best accuracy, but scales poorly as the number of exploratory behaviors and objects increases.

In this work, we investigate using exploratory behaviors to learn a novel predicate on a time budget without sacrificing grounding accuracy. We use two datasets of predicate-object relationships that include both visual ("red," "cylinder") and non-visual ("heavy," "full") predicates that require haptic and auditory feedback from an embodied robot. We compare methods for deciding which behaviors to perform when exploring objects in order to learn a new predicate, beyond the obvious time-consuming option of performing all behaviors. One possibility is to utilize unsupervised information in the form of word embeddings, such as those produced by Word2Vec (Mikolov et al. 2013). The distance between two words' embedding vectors suggests their semantic similarity. If "green" is close to "red" in the embedding space, a robot may be able to learn "green" using just the exploratory behaviors that determine whether an object is "red." Another possibility, if a robot is operating in a shared space with humans, is to ask a human which behaviors they would perform to evaluate the predicate.

We demonstrate that word embeddings help learn predicates using fewer behaviors. Our approach is independent of the embedding vectors, and we compare embeddings from two different corpora, noting that as the categorical quality of the embeddings improve (e.g. colors close to colors, weights close to weights), so should the gains achieved by our approach. We also show that using them together with human-provided behavior annotations speeds up learning in a domain of real-world objects with predicates from organic human descriptions in an embodied setting where behaviors must be performed in a certain order (e.g. an object must be *grasped* before it can be *lifted*).

Related Work

Most past research in language grounding has focused on using the visual sensory modality. Recent research has demonstrated that non-visual modalities can also be used to improve a robot's ability to ground semantic information (Araki et al. 2012; Chu et al. 2013; Silberer and Lapata 2014;

Kiela and Clark 2015; Thomason et al. 2016; Gao et al. 2016b; Alomari et al. 2017). For example, past work demonstrated that when humans teach a robot words that describe objects, learning performance is improved if the robot considers non-visual sensory information (e.g. audio and haptics) detected when manipulating the objects (Thomason et al. 2016).

A major limitation of these approaches is that they require the robot to perform exhaustive object exploration, i.e., the robot must explore each object with some fixed number of exploratory actions (e.g., grasp, lift, shake, push, etc.) during which it records non-visual sensory data. For example, in past work on learning multi-modal classifiers for a set of haptic adjectives, a robot performed seven different exploratory behaviors on 51 objects for a total of five times (Chu et al. 2013). While some methods have been proposed for how a robot should sequence its behaviors to minimize exploration time when classifying a novel object, these approaches still require that exhaustive exploration be performed during training (Sinapov et al. 2014; Zhang et al. 2017).

One possible way to address this problem is to estimate the relevance of each behavior for the task of learning a novel predicate or category. Sinapov *et al.* (Sinapov, Schenck, and Stoytchev 2014) show that a robot's learning performance on a novel predicate (e.g., "red") can be improved if the robot has some prior information about the predicate's similarity to known words (e.g., "blue" and "green"). The paper stops short of exploring where such a prior could originate. In this work, we answer this question by using word embeddings to estimate the relevance of known words to novel ones to be learned. Additionally, we gather annotations from humans about which behaviors they perceive as relevant for a given predicate (Figure 5). This is related to past human annotations gathered for relevant sensory modalities of words (Lynott and Connell 2009).

To guide the robot's exploration when learning a new word, we use distributional semantics to map words into high-dimensional vector spaces where their vector distances carry semantic information. Word2Vec uses a neural skip-gram model to create an embedding space for words given a large corpus (Mikolov et al. 2013). Related strategies consider context embeddings of words as well (Melamud, Levy, and Dagan 2015). Past work has created multi-modal Word2Vec-style embeddings that consider textual context together with visual (Silberer and Lapata 2014; Lazaridou, Pham, and Baroni 2015; Kottur et al. 2016) or audio (Vijayakumar, Vedantam, and Parikh 2017) context.

Recent work has used word embeddings to predict unseen verb causality information from seen verbs (Gao et al. 2016a), and unseen noun affordances from seen nouns (Fulda et al. 2017). These are similar in spirit to our use of unsupervised word embeddings created from large, unannotated text corpora to assist with a supervised grounded language learning problem—predicting the multi-modal representations most helpful for understanding a novel predicate.

The problem we address bears some similarities to the zero-shot learning problem (Xian, Schiele, and Akata 2017; Fu et al. 2015; Kodirov et al. 2015). In zero-shot learning, the task is to produce a classifier for a novel class label for which labeled data is unavailable, given some descriptor of that class label. In our case, the task for our robot is to produce a behavioral exploration policy when learning a new word given an embedding that relates the novel word to ones that are already learned. To our knowledge, this problem has not been addressed in the zero-shot learning literature.

Methodology

Let P be a set of predicates and O be a set of objects. Let the label function $\mathcal{L}(p, o) \in \{-1, 1\}$ indicate whether predicate $p \in P$ holds true for object $o \in O$. Let B be the set of available exploratory behaviors and let C be a set of sensorimotor contexts, such that each context corresponds to a combination of a behavior (e.g., grasping an object) and a sensory modality (e.g., auditory features extracted from the sound detected during grasping).

The robot's task is to learn predicate classification models that can predict whether or not a predicate holds true for an object given its multi-modal behavioral observations of that object. To learn such a model, in this work we adopt the method proposed in (Sinapov, Schenck, and Stoytchev 2014), in which the robot learns an individual grounding classifier $\mathrm{G}_{p,c}$ for each sensorimotor context c and predicate p. To determine whether the predicate applies to object $o \in O$, the weighted combination of these context-specific classifier outputs gives a consensus decision $d(p, o) \in \{-1, 1\}$. That is,

$$d(p, o) = sgn\left(\sum_{c \in C} w_{p,c} \mathrm{G}_{p,c}(o)\right), \quad (1)$$

where $w_{p,c}$ is the estimated reliability weight of context c for estimating whether predicate p applies to a given object.

One possible way of setting the weight $w_{p,c}$ is to make it proportional to the classification performance (e.g., Cohen's κ) of the classification function $\mathrm{G}_{p,c}$ as estimated from training data (Sinapov, Schenck, and Stoytchev 2014). Once these reliability weights have been estimated, we hypothesize that the robot can then perform only a subset of its behaviors to achieve high classification performance on novel objects. Below, we formulate the problem of how a robot can estimate surrogate reliability weights for a *novel* predicate for which *no* training data is available.

Problem Formulation. Given a set of known predicates P, their labels on a set of explored objects $O_E \subset O$, an unseen predicate q to be learned, and an unexplored set of objects $O_U \subset O$ (with $O_E \cap O_U = \emptyset$) labeled for predicate q, we explore strategies for learning a classifier for q on an exploration time budget without sacrificing accuracy.

The robot's task is two-fold: 1) estimate surrogate weights $w_{q,c}$ for the novel predicate q and each context $c \in C$; and 2) determine the order(s) in which to perform behaviors $b \in B$ given their cost (e.g., time) and the estimated weights associated with their contexts on O_U. Reliability weights for q can then be re-estimated at test time from the newly explored, labeled objects.

Estimating Unseen Predicate Context Reliability Weights. A baseline strategy for estimating surrogate $w_{q,c}$ is to assign a uniform weight per context. We can also use word embedding distances to share context weights from known predicates P to unknown predicate q. For every pair of predicates $p, q \in P$ with word embedding vectors v_p, v_q we calculate the similarity as the positive cosine distance:

$$poscos(p, q) = \frac{1}{2}(1 + cos(v_p, v_q)) \in [0, 1]. \quad (2)$$

It is common to use cosine distance in high-dimensional embedding spaces to measure word vector dissimilarities because it is independent of features' magnitudes. We find the top-k most lexically similar predicates to q in an embedding space, $P_q \subseteq P, |P_q| = k$ (allowing more than k in the event of a tie) and take a similarity-weighted average of $w_{p,c}$,

$$w_{q,c} \approx \frac{1}{|P_q|} \sum_{p \in P_q} poscos(p, q) w_{p,c}. \quad (3)$$

Expected Values for Behaviors. Given a weight for every context, we calculate weights $w_{q,b}$ at the behavior level. These are obtained by calculating training object decisions at the context level, aggregating them using weights $w_{q,c}$ (for c a context of behavior b), and calculating κ confidences based on those behavior-specific decisions across the training objects. Then we can calculate the expected value for each behavior as:

$$v(b) = w_{q,b} + \epsilon, \quad (4)$$

for some small ϵ such that behaviors with no confidence weight are not zero valued at training time (since they may yet prove useful at testing time).

Experiments

We performed experiments on two datasets. The first has a small number of predicates and a representative set of objects that readily support effective learning, and thus clearly demonstrates the utility of the proposed approach. The second dataset has a large number of predicates that arose organically during human-robot interaction for a diverse set of household objects, and thus learning the predicates is much more challenging.

Experiment 1: Learning object colors, weights, and contents

We demonstrate the effectiveness of surrogate reliability weight estimation using word embeddings to predict relevant contexts for a novel predicate given known predicates.

Dataset Description. We first used the dataset described by (Sinapov, Schenck, and Stoytchev 2014), in which a robot explored 36 different objects using 11 prototypical exploratory behaviors: *look*, *grasp*, *lift*, *shake*, *shake-fast*, *lower*, *drop*, *push*, *poke*, *tap*, and *press* to gather sensory information from: proprioceptive joint-torque sensors for all 7 joints, audio from an Audio-Technica U853AW cardioid microphone, and vision from a Microsoft Kinect sensor. The objects were identical containers except along 3 different attributes: 1) color: *red*, *green*, *blue*; 2) weight: *light*, *medium*, *heavy*; and 3) contents: *beans*, *rice*, *glass*, *screws*. These variations resulted in $3 \times 3 \times 4 = 12$ total predicates in the set P that the robot was tasked with learning.

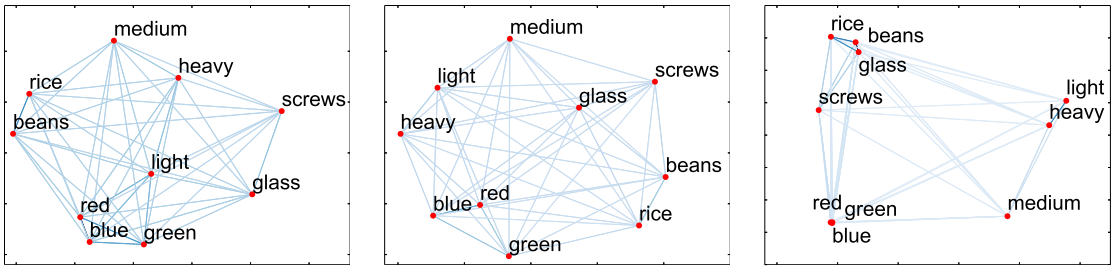
During the execution of the *look* behavior, the robot perceived 2 different sensory modalities, one corresponding to a color histogram of the object in the foreground, and the other comprising of a reduced size 10×10 RGB image of the object. For the remaining interactive behaviors, the robot recorded 2 types of sensory features, auditory and haptic, produced by the interaction with the objects. Thus, the robot's set of sensorimotor contexts was of size $|C| = 11 \times 2 = 22$.

Sample Predicate Embeddings. Figure 1 (a) shows a sample 2D projection of the Google News Word2Vec embeddings[1] corresponding to the 10 predicates in this dataset. Figure 1 (b) shows the 2D projection for the lexical substitution-focused embeddings that consider context embeddings (Melamud, Levy, and Dagan 2015). The projection was computed using Multi-Dimensional Scaling (MDS) (Kruskal and Wish 1978).

Figure 1 (c) shows an embedding based on each predicate's reliability weights estimated as agreement κ associated with each sensorimotor context in C. For each predicate $p \in P$, a feature vector f_p of size $|C|$ was computed such that the i^{th} entry corresponded to the confidence κ_i for context $c_i \in C$. These vectors were used to compute a $|P| \times |P|$ distance matrix using Euclidean distance. Notably, the visualizations show that there is some shared structure between the lexical embeddings and this sensorimotor embedding. In particular, attributes of similar types (e.g. colors) appear close together in both embedding types. We show that exploiting this shared structure can be used to improve learning novel predicates.

Evaluation and Results. The proposed methodology was evaluated using a "leave one predicate out" approach: during each run, the robot learned multi-modal grounded classifiers for 10 of the 11 total predicates P, using 12 fully explored and labeled objects that were randomly sampled from the entire set of 36 objects. When learning the remaining predicate, the robot was given a budget of N behaviors to use during both training and testing. The robot estimated the context reliability weights for the novel predicate using the lexical substitution-focused word embeddings (Melamud, Levy, and Dagan 2015) via Eq. 3, with $k = 7$, and propagated these to the behavior level. These estimates were then used to compute a distribution over behaviors B, which was used to sample a subset of size N (the budget) used for both

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



a) google news embedding b) lexsub embedding c) sensorimotor embedding

Figure 1: a) 2D projection of the Google News Word2Vec embedding of the 10 predicates used in the first experiment; b) 2D projection of the lexical substitution-focused embedding; c) 2D projection of an embedding constructed based on the relevant sensorimotor contexts for each of the 10 predicates. Shared structure can be seen between the word embeddings (a,b) and the sensorimotor embeddings of robot experience (c), which we leverage for learning novel predicates.

training and testing. In this experiment, we do not explicitly model behavior transitions, but instead assume that any behavior can be performed at any time and that all behaviors have equal cost. The context-specific predicate recognition models were implemented by a Support Vector Machine (SVM) with an RBF kernel.

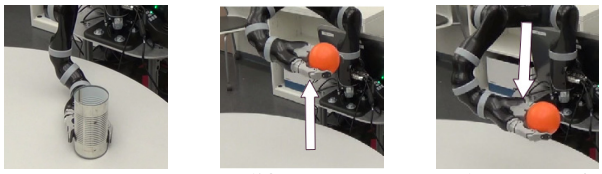
The results of this test are shown in Figure 2. Each of the three plots contains the average κ recognition rates for the three types of predicates. The proposed method is compared against the baseline approach of randomly selecting b behaviors using a uniform prior. Given sufficient budget, all methods perform all behaviors and achieve identical accuracy; examining these reduced budgets shows the effectiveness of our approach under exploration time constraints.

The proposed method enables the robot to reach good recognition rates ($\kappa > 0.95$) faster than random exploration, with the difference especially noticeable for color- and contents-related predicates. Figure 3 shows recognition results using a budget of $N = 1$ behavior for two different embeddings: lexical substitution and Google News. For some of the predicates, the lexical substitution embedding performs substantially better; in particular, the Google News embeddings links the word *light* with the colors and thus, the first behavior chosen when learning it tends to be *look*, which does not provide informative signals regarding the weight of the object (they all have the same size). On the other hand, the lexical substitution embedding puts *light* closer to the other two weight-related predicates and thus achieves the best performance.

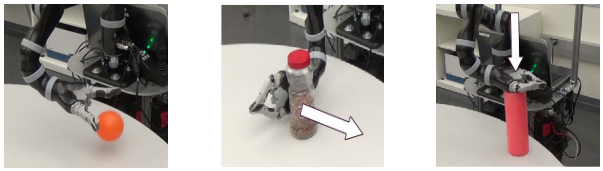
Below, we evaluate the proposed method on a much more challenging dataset in which the robot was tasked with learning words provided by everyday human users, and constrained to perform behaviors in a realistic order while considering the time it takes to perform each behavior.

Experiment 2: Learning words from everyday human users

We use feature representations from multiple behaviors and modalities for 32 objects using 8 exploratory behaviors



grasp (22s) lift (11.1s) lower (10.6s)

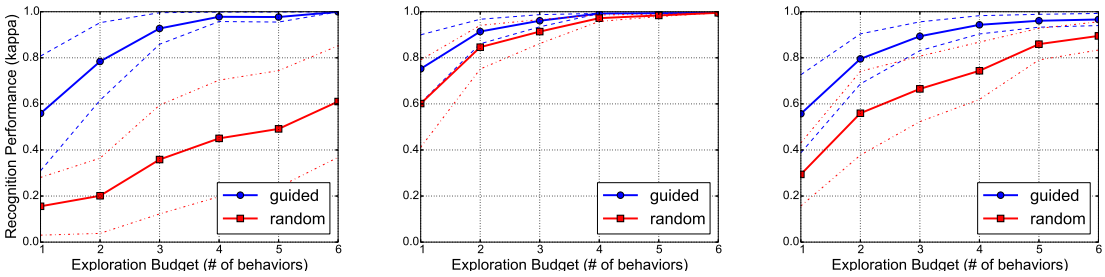


drop (9.8s) push (22s) press (22s)

Figure 4: The behaviors used to explore objects and the time in seconds for each. In addition, the *hold* (5.7s) behavior was performed by holding the object in place. The *look* (0.8s) behavior was also performed.

(Figure 4), collected by researchers for an object ordering task (Sinapov et al. 2016). For every object, there are features from every sensorimotor context. For the *grasp*, *lift*, *lower*, *drop*, *press*, *push*, and *hold* behaviors, *audio* (discrete fourier transform using 65 frequency bins) and *haptic* (joint efforts and joint positions for 6 joints) information is available. For the *look* behavior, *color* (RGB color histogram using 8 bins per channel), *shape* (fast point feature histogram (fpfh)), and *deep* (VGG (Simonyan and Zisserman 2014)) features are available (the latter drawn from the features added by (Thomason et al. 2016) for language grounding). These modalities result in $|C| = 7 \times 2 + 1 \times 3 = 17$ contexts.

Predicate Annotations. We consider 81 predicates available from a human-robot interaction dataset in which humans gave unrestricted natural language descriptions of objects (Thomason et al. 2016). In this work, we gathered full annotations between those 81 predicates and the 32 objects in (Sinapov et al. 2016) (defining $\mathcal{L}(p, o)$ for every predicate p and object o). We gathered 3 annotators' opinions about



a) *Color* Predicates b) *Weight* Predicates c) *Contents* Predicates

Figure 2: Test-time κ performance of classifiers for learning a new predicate based on the reliability weight estimation strategy used at test time as more behaviors are allowed. **random** chooses the next exploratory behavior at random, while **guided** uses word embeddings to select known neighbor predicates from which to estimate reliability weights for behaviors. The dotted lines denote the variance over 75 simulation runs.

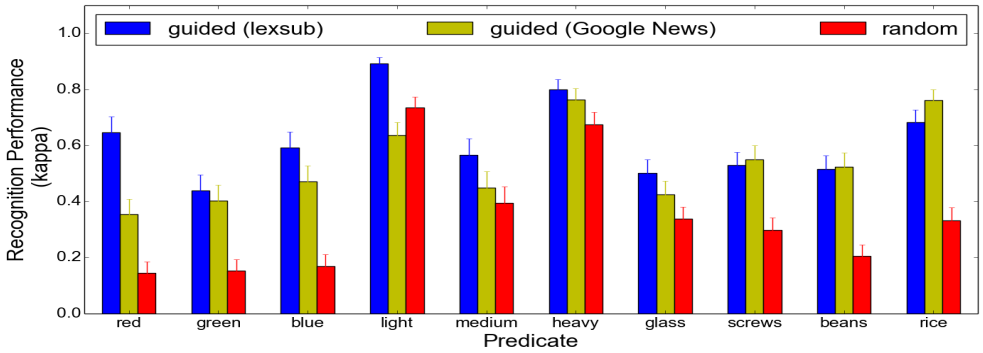


Figure 3: Recognition results for all 10 predicates using just 1 exploratory behavior selected according to three different conditions: *guided* with lexical substitution predicate embedding, *guided* with Google News predicate embedding, and *random*. The bars denote standard error.

whether each predicate applied to each object. We took a majority vote between the 3 annotators when there was a disagreement. To reduce annotator fatigue, each annotator labeled predicates for 8 of the 32 total objects, requiring 12 annotators in total to gather labels. The average pairwise κ agreement between annotators was 0.576 (reasonable agreement). Figure 5 shows all the predicates given positive labels for a sample object.

Behavior Annotations. For each of the 81 predicates, we gathered annotations in order to create a distribution over behaviors relevant for that predicate. Annotators were asked to mark which exploratory behaviors 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 pairwise κ agreement with all other annotators was less than 0.4. This cutoff left us with 8 annotators whose average agreement was

text, bright, cup, large, round, heavy, container, red, full, water, cylindrical, colored, thing, hollow, top, plastic, white, cap, cylinder, medium-sized, tall, liquid, object, bottle

Figure 5: Predicates with positive labels for the object in the picture, from annotations gathered in this work.

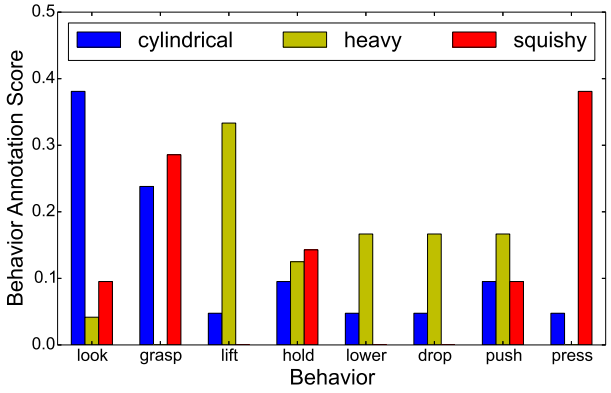


Figure 6: Behavior annotations for three predicates in the dataset: "cylindrical", "heavy" and "squishy". Scores correspond to the number of times annotators rated the behavior as relevant for recognizing whether the predicate applied to an arbitrary object.

$\kappa = .475$. We assign each behavior a value for each predicate of the ratio of annotators (out of these 8) who marked it relevant, so that for every $p \in P, b \in B$ we have an annotation score $A(p, b) \in [0, 1]$. Figure 6 shows the behavior annotation scores for three predicates. We release the predicate-object labels and predicate behavior annotations as a supplementary dataset.[2]

In addition to estimating $w_{q,c}$ from Eq. 3 (e.g. top-k nearest word embedding lexical neighbor predicates), we estimate it from behavior annotations alone (Eq. 5) and from an interpolation of behavior annotations those semantic neighbors (Eq. 6). For C_b the set of contexts for behavior b and b_c the behavior associated with context c:

$$w_{q,c} \approx \frac{1}{|C_b|} A(q, b_c); \tag{5}$$

$$w_{q,c} \approx \frac{1}{|C_b|} A(q, b_c) * \frac{1}{|P_q|} \sum_{p \in P_q} poscos(p, q) w_{p,c}. \tag{6}$$

Choosing an Exploration Policy. Given the values of each behavior (Eq. 4) for an unknown predicate q, the presuppositions of each behavior, the time to perform each behavior $t(b)$, and a time limit per object for exploration T, we can sample a sequence of behaviors to use when evaluating predicate q. Figure 7 describes the effects of behaviors on the object being explored, while Figure 4 gives the time in seconds to perform each. Because there are 5 observations per behavior per object available, each behavior can be performed in an exploration policy up to five times, making enumerating all policies intractable.

We take a Monte-Carlo-style approach, sampling a large number of behavior sequences through weighted random walks, then choosing one sequence among those that maximize reliability weight while minimizing time. To sample a sequence of behaviors, we start at the "on table" state (Figure 7), choosing any available behavior with probability proportional to $v(b)$ (Eq. 4) with respect to other available behaviors. For example, from the "on table" state, the probability of choosing *press* is

$$p(\text{press}) = \frac{v(\text{press})}{v(\text{look}) + v(\text{press}) + v(\text{push}) + v(\text{grasp})},$$

assuming *press*, *look*, *push*, and *grasp* have each been performed fewer than 5 times and there is enough remaining time in the budget given the sequence so far to execute each alone (e.g. $t(\text{press}) \leq T$). A sampled sequence ends when these constraints are met by no outgoing behaviors.

In our experiments, we sample 100 sequences S for every training trial. Of those sampled, we first select the subset $\hat{S}$ of sequences with the highest value, then randomly choose one with the shortest exploration time,

$$\hat{S} = \text{set-argmax}_{s \in S} \left(\sum_{b \in s} v(b) \right);$$

$$s^* \in \text{argmin}_{S \in \hat{S}} \left(\sum_{b \in S} t(b) \right).$$

The chosen sequence s^* is used to explore the unseen objects O_U, extracting features for training classifiers for predicate q.

Experiments and Results We randomly split the 32 objects into 16 explored objects O_E and 16 unexplored objects O_U. We then perform leave-one-predicate-out cross validation, holding predicate q out. For predicate q, we are given the labels $\mathcal{L}(q, o)$ for $o \in O_U$. We then perform leave-one-object-out cross validation, deciding on a training behavior sequence for q, using it to explore 15 of the unexplored objects, re-estimating context reliability weights as κ agreement, and finally exploring the held-out object and assigning a label for q based on these new reliability weights. In this way, we can obtain agreement statistics with true labels for every held-out predicate, aggregating these to compare different surrogate reliability weight estimations for choosing an exploration policy.

Our leave-one-predicate-out experiment operates over the 48 predicates for which O_U had at least 2 positive and 2 negative object examples for the predicate. We calculate word embedding distance (Eq. 2) using Google News Word2Vec embeddings,[3] use linear SVMs as context-level classifiers, set $k = 3$ (Eq. 3), and set $\epsilon = 0.001$ (Eq. 4). For every time budget T and surrogate reliability estimate compared, the behavior sequence sampling and leave-one-object-out cross validation was repeated 100 times to get average performance. Figure 8 shows these average performances. The time budgets are chosen so that each behavior has time to be performed one, two, and three times each, if the policy chooses homogeneously.[4]

[2] [[URL redacted for anonymous review]]

[3] The alternative, lexical substitution-focused embeddings (Mohan, Mininger, and Laird 2013) perform similarly.

[4] With sufficient time, all methods are able to perform all behaviors five times (the maximum), achieving convergent performance.

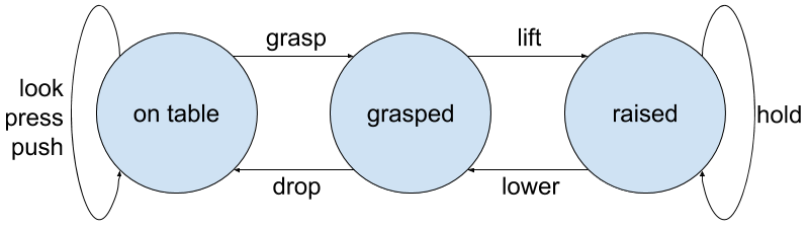


Figure 7: Exploratory behavior actions as transitions in an object state graph.

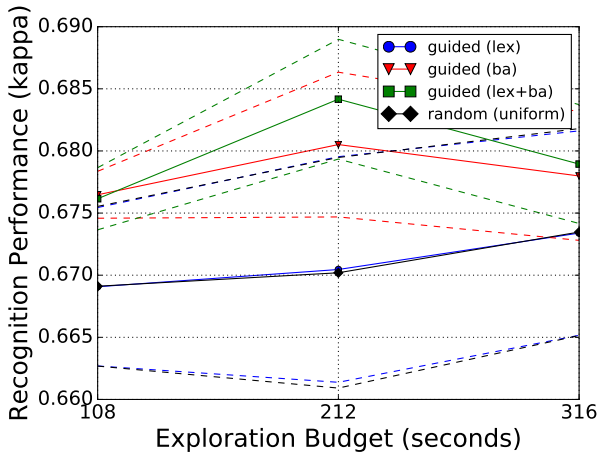


Figure 8: Test-time κ performance of classifiers for learning a new predicate based on reliability weight estimation strategy used at test time for three time budgets. **uniform** assigns reliability $\frac{1}{|C|}$ to each context, **lex** (Eq. 3) estimates reliability weights from neighbor predicates, **ba** (Eq. 5) from behavior annotations alone, and **ba+lex** (Eq. 6) from behavior annotations interpolated with **lex**, respectively. The dotted lines denote the variance over 100 simulation runs. Across all predicates, **lex** alone does not outperform the **uniform** baseline, but when combined with behavior annotations **ba+lex** achieves the best performance overall.

Figure 8 shows the average κ agreement achieved by grounding classifiers trained under different surrogate weight estimation strategies. In this more difficult set of objects and predicates, borrowing weights from nearest lexical neighbors in word embedding space (**lex**) is insufficient to improve grounding accuracy on a behavior time budget. Unlike the clear-cut predicates of Experiment 1, the predicates arising from human users in this dataset do not form as clearly defined semantic clusters as those visible in Figure 1.

However, behavior annotations (**ba**) improves performance, and the best performance for grounding classifiers is achieved when considering these together with lexical neighbor information (**ba+lex**). We postulate that this occurs because there is a slight mismatch between the behaviors that humans would use to determine properties versus what is actually helpful to a robot. Conversely, human intuitions about which behaviors are relevant help prune out information from erroneous lexical neighbors in this more complicated set of predicates.

These results demonstrate that gathering behavior annotations for an unseen predicate can improve grounding performance on a time budget, and performance is further boosted by using word embeddings to share neighboring predicates' reliability weights.

Conclusions

Current methods for grounding object concepts in behavioral exploration and multi-modal perception suffer from the limitation that a robot needs to exhaustively perform all of its actions to figure out which ones are useful for learning the target concept. To address this problem, this paper proposed a framework for guiding a robot's behavioral exploration of objects when learning new words. In the proposed framework, given a novel word, the robot computes an exploration policy specific to that word by relating it via word embeddings to words that have already been learned.

Our first experiment demonstrated that our method allows the robot to learn new words faster, in terms of the number of different behaviors the robot needs to perform on objects to learn the target word. In our second experiment, we also demonstrated that behavior annotations gathered from human users can be integrated into the framework to further improve predicate recognition performance under a time budget as well as physical and temporal constraints.

In future work, behavior annotations could be gathered from human users on-the-fly in an embodied dialog setting, using a learned human-robot dialog policy to know when behavior annotation questions are warranted. Using modality annotations (Lynott and Connell 2009) may further boost performance. Additionally, rather than deciding on a static exploration policy, in the future we would like to determine an exploration policy dynamically, given that some behaviors can change the state of the object unexpectedly (e.g. if *push* knocks an object off a table). In Experiment 1, we compared different word embeddings and their relationship to sensorimotor embeddings; in the future, it may be worthwhile to store multiple word embedding options and learn which embedding space to use given a novel predicate.

Acknowledgments

This work is supported by a National Science Foundation Graduate Research Fellowship to the first author 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 (IIS-1637736, IIS-1651089, IIS-1724157), Intel, Raytheon, and Lockheed Martin. Peter Stone serves on the Board of Directors of Cogitai, Inc. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

References

Alomari, M.; Duckworth, P.; Hogg, D. C.; and Cohn, A. G. 2017. Natural language acquisition and grounding for embodied robotic systems. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4349–4356.

Araki, T.; Nakamura, T.; Nagai, T.; Funakoshi, K.; Nakano, M.; and Iwahashi, N. 2012. Online object categorization using multimodal information autonomously acquired by a mobile robot. *Advanced Robotics* 26(17):1995–2020.

Chu, V.; McMahon, I.; Riano, L.; McDonald, C. G.; He, Q.; Perez-Tejada, J. M.; Arrigo, M.; Fitter, N.; Nappo, J. C.; Darrell, T.; et al. 2013. Using robotic exploratory procedures to learn the meaning of haptic adjectives. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 3048–3055. IEEE.

Fu, Z.; Xiang, T.; Kodirov, E.; and Gong, S. 2015. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2635–2644.

Fulda, N.; Ricks, D.; Murdoch, B.; and Wingate, D. 2017. What can you do with a rock? affordance extraction via word embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, 1039–1045.

Gao, Q.; Doering, M.; Yang, S.; and Chai, J. Y. 2016a. Physical causality of action verbs in grounded language understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gao, Y.; Hendricks, L. A.; Kuchenbecker, K. J.; and Darrell, T. 2016b. Deep learning for tactile understanding from visual and haptic data. In *International Conference on Robotics and Automation (ICRA)*, 536–543. IEEE.

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

Kiela, D., and Clark, S. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2461–2470.

Kodirov, E.; Xiang, T.; Fu, Z.; and Gong, S. 2015. Unsupervised domain adaptation for zero-shot learning. In *The IEEE International Conference on Computer Vision (ICCV)*.

Kottur, S.; Vedantam, R.; Moura, J. M. F.; and Parikh, D. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Kruskal, J. B., and Wish, M. 1978. *Multidimensional scaling*, volume 11. Sage.

Lazaridou, A.; Pham, N. T.; and Baroni, M. 2015. Combining language and vision with a multimodal skipgram model. In *The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

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

Melamud, O.; Levy, O.; and Dagan, I. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 1–7. Denver, Colorado: Association for Computational Linguistics.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, 3111–3119.

Mohan, S.; Mininger, A. H.; and Laird, J. E. 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*.

Silberer, C., and Lapata, M. 2014. Grounded meaning representations with autoencoders. In *In Proceedings of the Association for Computational Linguistics (ACL)*.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *Computing Research Repository* abs/1409.1556.

Sinapov, J.; Schenck, C.; Staley, K.; Sukhoy, V.; and Stoytchev, A. 2014. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems* 62(5):632–645.

Sinapov, J.; Khante, P.; Svetlik, M.; and Stone, P. 2016. Learning to order objects using haptic and proprioceptive exploratory behaviors. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*.

Sinapov, J.; Schenck, C.; and Stoytchev, A. 2014. Learning relational object categories using behavioral exploration and multimodal perception. In *IEEE International Conference on Robotics and Automation*.

Thomason, J.; Sinapov, J.; Svetlik, M.; Stone, P.; and Mooney, R. 2016. Learning multi-modal grounded linguistic semantics by playing "I spy". In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 3477–3483.

Vijayakumar, A. K.; Vedantam, R.; and Parikh, D. 2017. Sound-word2vec: Learning word representations grounded in sounds. *arXiv preprint arXiv:1703.01720*.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning - the good, the bad and the ugly. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR-17)*.

Zhang, S.; Sinapov, J.; Wei, S.; and Stone, P. 2017. Robot behavioral exploration and multimodal perception using pomdps. In *AAAI Spring Symposium on Interactive Multisensory Object Perception for Embodied Agents*.