# Grounding Natural Language Instructions with Unknown Object References using Learned Visual Attributes

**Nicole Glabinski** and **Rohan Paul** and **Nicholas Roy***
Massachusetts Institute of Technology, USA

## Introduction

Our goal is to enable robots to interpret or *ground* natural language instructions from a human in the context of the operating workspace. We address the problem of grounding linguistic references within an instruction associated with semantic objects populating the scene. In realistic workspaces, the types and locations of objects may only be partially known to the robot. Hence, the robot may encounter novel objects types for which pre-trained visual classifiers may not exist and hence cannot be recognized from an image. Further, ambiguity arises when an instruction may refer to one of several unknown object types detected in the scene.

Recent progress has been made in probabilistic approaches that relate language with semantic concepts derived from the environment, such as (Tellex et al. 2011), (Howard, Tellex, and Roy 2014), (Matuszek, Fox, and Koscher 2010), (Chai et al. 2016), (Paul et al. 2016), (Mei, Bansal, and Walter 2016) and (Shridhar and Hsu 2017). These models rely on an *a priori* known set of object types and hence cannot ground references to novel objects in the scene. Complementary work exists in the computer vision community, where researches explore zero shot learning, i.e., the ability to learn a classifier for unseen or novel objects. Approaches in this category learn a classification model based on primitive visual attribute such as shape, texture, patterns etc. Examples include (Ferrari and Zisserman 2007), (Farhadi et al. 2009), (Lampert, Nickisch, and Harmeling 2009), and (Jayaraman and Grauman 2014).

In this work, we consider the scenario of grounding unknown noun references in scenes populated multiple objects that the robot cannot recognize. We build on prior work and present a model that resolves ambiguity amongst unknown groundings by eliciting visual attribute descriptions from the human. Linguistic attribute descriptions that convey texture, material, shape, and appearance is fused with a probabilistic language grounding framework by introducing probabilistic factors that predict a set of visual attributes from an image. The attribute predictor is trained from a crowd sourced dense image annotation data set. Figure 1 presents an overview of the approach.

*Authors gratefully acknowledge support from the Robotics Collaborative Technology Alliance (RCTA) of the US Army, Toyota Research Institute and the National Science Foundation.

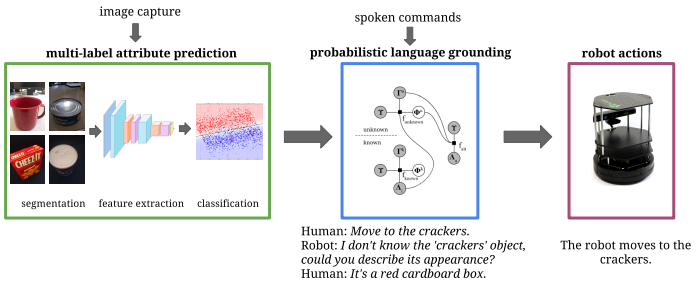



Figure 1: Overview. An image attribute prediction model is combined with a language grounding model to resolve ambiguity among multiple grounding candidates for an unknown object reference by eliciting descriptive attributes from the human.

## Technical Approach

We first describe the visual attribute prediction model and then discuss the language grounding model that enables the robot to interpret unknown object references using the learned visual attributes.

### Multi-label Attribute Prediction

Given an image of an object extracted from the scene using object segmentation, we seek a model that estimates how well each attribute describes the object. The attribute predictor produces probabilities for each of the fixed attributes in Table 1. We collected training data from the Visual Genome dense image annotation dataset (Krishna et al. 2016), that contains human-generated descriptions and bounding boxes of individual objects within images. We extracted matching word annotations or synonymous for an attribute as wells as the tight bounding box labeled with the attribute. Following Sharif et al. (Sharif Razavian et al. 2014), we use learned CaffeNet features from the last layer of a CaffeNet network (Jia et al. 2014) pre-trained for object recognition on the ImageNet dataset. These features are used for training the one-vs-all logistic regression for multi-label classification.

| | |
|---|---|
| **Color** | red, blue, green, yellow, orange, purple, brown, black, white |
| **Shape** | square, rectangular, cylindrical, round, flat, curved |
| **Material** | cardboard, metallic, cloth, glass, plastic, wooden |

Table 1: Visual attributes learned by the model.

## Grounding Instructions with Unknown Objects

We build on the Distributed Correspondence Graph (DCG) model (Howard, Tellex, and Roy 2014) that associates an input instruction $\Lambda$ with semantic concepts or *groundings* $\Gamma$ derived from the robot's world model. Groundings include objects, spatial regions as well as goals specifying future actions that the robot can take. The grounding process is mediated by a set of binary correspondences $\Phi$ that express the degree to which a phrase in the input instruction relates to a candidate grounding. Hence, the grounding problem can be posed as determining the most probable correspondences:

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}}\, P(\Phi|\Gamma, \Lambda). \tag{1}$$

The world model consists of a set of detected objects with their location and type as classified by the perception system. Following (Tucker et al. 2017), an object is assigned the *unknown* type in case the classification likelihood is uninformed (characterized using a margin or entropy based measure). The space of groundings $\Gamma$ is augmented with symbols associated with *unknown* object detections $\Gamma_u$ in addition to grounding variables associated with known object types $\Gamma_k$. Assuming grounding variables to be independent given the input instruction, the joint likelihood can be expressed as:

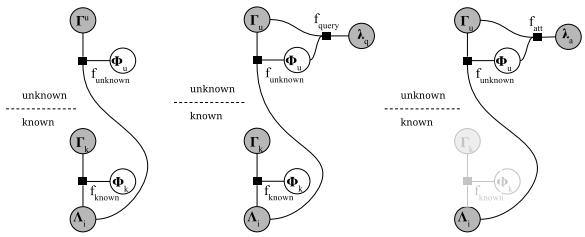
$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \overbrace{P(\Phi_k|\Lambda, \Gamma_k)}^{f_{known}} \overbrace{P(\Phi_u|\Lambda, \Gamma_u)}^{f_{unknown}}. \tag{2}$$

Figure 2a illustrates the graphical model. The equation factorizes further over phrases according to the parse structure of the input instruction. Each factor is realized as a log-linear model and trained using a language corpus aligned with simulated scenes. The model infers unknown groundings when confronted with unknown phrase references and uses the linguistic reference paired with observations to learn a grounding model for the new object for future inferences. However, the model cannot resolve ambiguity in case there are multiple unknown objects in the scene. We incorporate attribute elicitation in the model by generating language query $\lambda_q$ conditioned on the set of unknown grounding variables expressed as the factor $f_{query}$. The language query requests for additional visual attributes describing the intended object incase there are multiple expressed unknown groundings $\Gamma_u$, see Figure 2b. Phrases in the response labeled as adjective (though part-of-speech tagging) constitute the observed language variable $\lambda_a$ and is introduced in the model in the factor $f_{att}$ shared with unknown grounding variables, Figure 2c. This factor models the likelihood of the visual attributes given the visual detection and is realized using the multi-label attribute prediction model discussed previously.

The inclusion of the attribute prediction factor results in an augmented model where the probable set of correspondences are estimated as:

$$\Phi^* = \operatorname{argmax}_{\Phi} \overbrace{P(\Phi_k|\Lambda_i, \Gamma_k)}^{f_{known}} \overbrace{P(\Phi_u|\Lambda_i, \Gamma_u)}^{f_{unknown}} \overbrace{P(\Phi_u|\lambda_a, \Gamma_u)}^{f_{att}}. \tag{3}$$

The additional descriptive attributes elicited by the human enable the model to resolve ambiguity among multiple unknown grounding candidates.

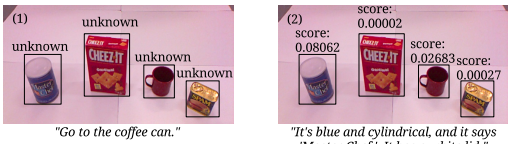



(a) (b) (c)

Figure 2: Graphical models instantiated during the three inference steps. (a) Grounding the initial instruction. An unknown grounding is inferred for an unknown object reference in the input instruction. (b) The factor $f_{query}$ determines whether to query. (c) If query was asked, response adjectives $\lambda_a$ added an incorporated as a factor associated with unknown groundings. Groundings with false correspondences (greyed out) are removed from the model at this stage.

## Preliminary Evaluation and Expected Results

Figure 3 presents a preliminary demonstration of the proposed model. using a Turtlebot mobile robot. The scene was populated with objects from the YCB object dataset (Calli et al. 2015) and imaged using a Kinect V1 camera. The Edge-Boxes (Zitnick and Dollar 2014) toolbox was used for object segmentation. The robot was presented with an instruction to move towards an object in the scene. All objects are unknown to the robot. The visual attributes described by the human enable the robot to determine the correct grounding as the coffee can among the multiple unknown groundings.




"Go to the coffee can." "It's blue and cylindrical, and it says 'Master Chef.' It has a white lid."

Figure 3: Preliminary demonstration. The correct grounding (coffee can) is determined by incorporating the likelihood of visual attributes elicited from the human.

Figure 4 presents the quantitative evaluation of the attribute prediction model. Attributes such as color and texture showed higher prediction accuracy. However, attributes such as flat or cloth-like showed lower accuracies.

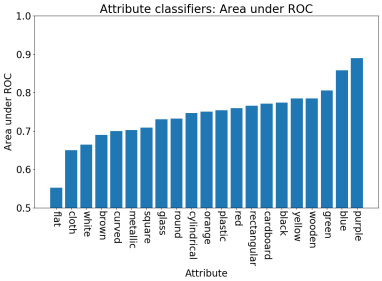



Figure 4: Area under ROC curve for attribute prediction model on the withheld Visual Genome data set (80:20 test train split).

Our plan for future evaluation includes evaluation on a larger data set, augmenting attribute prediction model trained on Visual Genome data with imagery from realistic robot workspaces and incorporating additional attributes such as text and spatial context in the scene.

## References


Calli, B.; Singh, A.; Walsman, A.; Srinivasa, S.; Abbeel, P.; and Dollar, A. 2015. The ycb object and model set: Towards common benchmarks for manipulation research. In *IEEE International Conference on Advanced Robotics (ICAR)*.

Chai, J. Y.; Fang, R.; Liu, C.; and She, L. 2016. Collaborative language grounding toward situated human-robot dialogue. *AI Magazine* 37(4).

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR*.

Ferrari, V., and Zisserman, A. 2007. Learning visual attributes. In *Advances in Neural Information Processing Systems*.

Howard, T. M.; Tellex, S.; and Roy, N. 2014. A natural language planner interface for mobile manipulators. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 6652–6659. IEEE.

Jayaraman, D., and Grauman, K. 2014. Zero-shot recognition with unreliable attributes. In *Advances in Neural Information Processing Systems*, 3464–3472.

Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; and Darrell, T. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678. ACM.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between class attribute transfer. In *Proceeding of Computer Vision and Pattern Recognition (CVPR)*.

Matuszek, C.; Fox, D.; and Koscher, K. 2010. Following directions using statistical machine translation. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, 251–258. IEEE.

Mei, H.; Bansal, M.; and Walter, M. R. 2016. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences.

Paul, R.; Arkin, J.; Roy, N.; and Howard, T. M. 2016. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *Proceedings of Robotics: Science and Systems*.

Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; and Carlsson, S. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 806–813.

Shridhar, M., and Hsu, D. 2017. Grounding spatio-semantic referring expressions for human-robot interaction. In *Proceedings of Robotics: Science and Systems Workshop on Spatial-Semantic Representations in Robotics*.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A. G.; Teller, S. J.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*.

Tucker, M.; Aksaray, D.; Paul, R.; Stein, G. J.; and Roy, N. 2017. Learning unknown groundings for natural language interaction with mobile robots. In *To appear in the International Symposium on Robotics Research (ISRR)*.

Zitnick, L., and Dollar, P. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*. European Conference on Computer Vision.