Improving Grounded Language Acquisition Efficiency Using Interactive Labeling

Nisha Pillai, Karan K. Budhraja, Cynthia Matuszek
npillai1 | karanb1 | cmat @ umbc.edu
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County, Baltimore, Maryland 21250

***Abstract*—Natural language has emerged as a powerful, intuitive interface for robot-human communication. There has been substantial work in recent years on *grounded language acquisition*, in which paired language and sensor data are used to create a model of how linguistic constructs apply to the perceivable world. While powerful, this approach suffers from the need for extensive natural language annotations. In this paper, we describe an initial pilot of a system that uses active learning to solicit annotations from a human interlocutor. Our results suggest that using active learning reduces the number of annotations necessary to learn such groundings, providing a strong justification for building a more robust version of such a system, and suggest some insights into human requirements for usability.**

I. Introduction

As robots have become safer and more capable, the idea of deploying them in situations where they interact with non-specialists (e.g., in homes, hospitals, or schools) has become more realistic. Natural language is an intuitive and widely understood way of conveying instructions and information. However, building appropriate language models for the wide range of real world situations and users is an enormous challenge, particularly in the area of *grounded language*, where language refers to objects and actions in a particular robot's perceptual world. In order to address this, it has become increasingly apparent that an interactive learning system where robots learn from unstructured communication with untrained users is necessary.

In previous work, we have demonstrated using unconstrained language paired with sensor and actuator data to support learning about novel object attributes (e.g., attributes that had no representation in the underlying language model until encountered) [12, 13]. While effective, these efforts depended on large amounts of user annotation—probably more than could be comfortably supplied by a single user. In order to accomplish the longer-term goal of learning groundings for descriptions of objects from end users, it is necessary to (1) reduce the amount of annotation necessary to learn about objects and (2) better understand the constraints on obtaining natural language descriptions from users.

We treat language descriptions as labels for novel visual percepts [12]. This pilot expands on our previous work in two ways: first, language is obtained from direct, real-time interaction with the user, rather than using a corpus of labels collected beforehand; second, the robot uses the performance of the visual classifiers to query the user proactively about unclear terms. In this way, the users (annotators) serve as oracles who provide labeled data. We present data about their experiences using the system, both with and without interactive learning.

The problem of learning the connection between novel percepts and novel language has been explored before. In this paper, we focus on the possible contribution of active learning to that problem, rather than difficulty of the vision problem. Our results show that interactive labeling reduces labeling costs and increases user acceptance compared to unguided labeling, and support the idea that actively seeking language labels is an important area of future work for robot language acquisition.

II. Related Work

Active learning, in which a learning system chooses data points to label, can allow for more efficient learning and better performance as well as more natural interaction [4, 8]. Using active learning in human-robot interactions presents its own challenges, raising issues not only of what questions to ask, but when and how to ask them [2, 18, 21].

In robotics, active learning is often encountered in learning from demonstration, in which a human tutor provides an example of an action [3], or in cases where the learning is used to indirectly support interaction [7, 9]. Our work is more similar to Kulick et al. [10] in that the active learning is used to directly guide the acquisition of language; however, we refer to different world objects and do not limit the system to yes/no questions.

The language learning component of this work fits into the category of *grounded language acquisition*, the integrated learning of language and environment [14]. This approach has led to successes, especially in using weakly supervised learning to support situated language learning over large corpora of data [20]. 'Environment' can take many forms, from implicit meaning in navigation instructions [6], to searching on [19] linguistic descriptions of spatial elements in video clips.

This work is an extension of Matuszek* et al. [12], in which a joint model of language and perception is used to acquire groundings for language describing the perceived characteristics of objects, allowing for learning words that have no pre-existing counterparts in an underlying formal grammar. However, we use a simpler bag-of-words model, which has been shown to be effective for this problem [20, 13].

III. Approach

A joint model of language and perception is used to acquire groundings for language describing the characteristics of objects in the environment. When new words are encountered, visual classifiers are created and trained on the perceptual context; for example, the classifier associated with "red" is trained as the word is encountered in conjunction with objects. We improve the efficiency of that process by using active learning [5], letting the robot preferentially seek data points that reduce uncertainty.

To evaluate efficiency and usability, a small number of people were asked to teach the robot about colors, either with or without active questioning about labels (see Figure 1). We describe these experiments and present preliminary results.

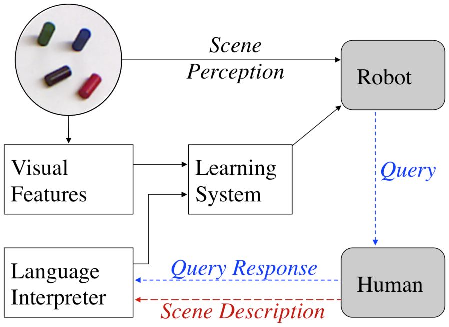

Fig. 1. The learning cycle. In the manual labeling case, the annotator is always asked to describe objects (red dashed line); when performing interactive labeling, the system prompts the annotator to either describe the object, or—in the case of a sufficiently trained classifier—to briefly verify the object's classification (blue dotted lines).

A. System

We use the widely-available Microsoft Kinect v2 RGB-D sensor, which is part of a custom mobile manipulation platform. For this work, only the sensors and speakers are used. When an object is placed in the robot's workspace, RGB-D information from the sensor is input to the learning system (as in Figure 1). The learning system uses this information to update its knowledge of objects. Such learning occurs iteratively for several objects.

We conducted a pilot with ten participants, who were asked to describe a series of 20 objects, divided into two cases (see Figure 1). In the first, which we call *manual annotation* or *manual labeling*, participants were asked to describe every object. In the second, *interactive labeling*, the robot tried to classify the objects; depending on the confidence of classification, the user either provided a description or was asked a *verification* question ("Is this an $\langle$*attribute*$\rangle$ object?", to which the user answered yes or no).

This verification question is much lower overhead than requesting a label, while still allowing the user to correct system errors. In future, we expect trained classifiers to have sufficient confidence to do away with verification questions entirely.

B. Classifier Learning

An RGB-D point cloud is extracted from the Kinect v2 sensor based on the scene. The object is then separated from the background using k-means clustering [11] on the depth data, and corresponding visual features—in this case, RGB values—are extracted. All processing is done using the Point Cloud Library (PCL) [17] in ROS [15]. For color classification, kernel descriptors [1] corresponding to color attributes are used. We use a bag-of-words model to find appropriate linguistic tokens, using keywords extracted using the Rapid Automatic Keyword Extraction (RAKE) [16] library.

A group of classifiers (one per keyword) is used by the system. Each classifier uses logistic regression with three inputs corresponding to RGB values. A threshold of 0.7 is used for the output of a classifier to describe a positive or negative sample, chosen empirically based on early studies.

Learning proceeds by periodically generating and training the classifiers, and updates the parameters for language and vision. A new classifier is created on encountering a previously-unseen keyword, and trained using the object in view as a positive datum; further objects described with that keyword serve as additional positive samples.

C. Grounded Language Acquisition

A joint model of language and perception is used to acquire groundings for language describing the characteristics of objects [12]. Upon encountering an unfamiliar language token, the system creates visual classifiers trained on percepts (for the vision system) and associated with tokens (for the language system)—in this case, keywords. Visual classifiers and language learning are treated as a single joint model with a shared learning objective. The result is a set of visual attribute classifiers that identify objects in the scene referred to in language.

As training data is added, the joint model obtains an expanding group of classifiers. Repetition of object attributes reinforces classifiers related to a corresponding keyword. In this way, keywords that actually refer to attributes of an object are likely to collect positive examples consisting of similar percepts (similar colors, in this example). Such classifiers (*e.g.*, associated with the word "red") will therefore gain predictive power. Classifiers with good predictive power then collectively form the grounded language model, while classifiers with reduced predictive power (*e.g.*, associated with "its") can eventually be pruned.

The model is *joint* in the sense that each classifier, when created, is associated with a language token, for example, `new-classifier-called-``red''`. This is a deliberately simple classification problem, as the goal of this work is to determine the possible positive contribution of active learning, rather than to solve a novel vision problem.

D. Interactive vs Manual Labeling

Our ultimate goal is to incorporate active learning into the joint learning model described, to reduce manual labeling and to gain information about the desired characteristics of an interactive learning interface. We are using a simple model of information gain: choosing an object at random (by the human annotator) and, if it cannot be classified accurately by existing trained classifiers, request a description.

When an object is chosen, the model attempts to classify its visual attributes with existing classifiers; if any of the classifiers gives a probability above a threshold, the system prompts the oracle (annotator) for confirmation and moves on, although it is flexible enough to accept any additional description about that object the oracle might wish to give at this point. If the object is new to the system, none of the classifiers indicate a probability above the threshold, or the system's classification is not confirmed, the human annotator would be asked to provide a description, which is immediately added to the model. The learning system uses logistic regression to calculate confidences.

IV. Learning Performance

In this section, we report on the performance of the trained model using manual vs. interactive labeling. While small sample sizes prevent this from being definitive, we find that interactive labeling—in which an annotator is asked to provide descriptions only for objects that cannot already

	"arc"	*"banana"*	*"blue"*	*"bottom"*	*"cylinder"*	*"green"*	*"half"*	*"object"*	*"rectangle"*	*"red"*	*"section"*	*"thin"*	*"triangle"*	*"yellow"*
green	0.025	0.002	0.252	0.024	0.182	**0.970**	0.019	0.195	0.019	0.000	0.055	0.017	0.079	0.022
red	0.024	0.005	0.007	0.003	0.124	0.000	0.116	0.250	0.030	**0.946**	0.041	0.031	0.072	0.024
yellow	0.008	0.048	0.024	0.010	0.086	0.099	0.029	0.210	0.046	0.004	0.063	0.010	0.010	**0.740**
blue	0.034	0.002	**0.628**	0.012	0.151	0.028	0.027	0.201	0.021	0.006	0.053	0.020	0.084	0.022

Fig. 2. Associations between classifiers created for keywords (top row) and ground truth colors (leftmost column). Only a small subset of representative classifiers are shown, since one is created for each keyword in the corpus. The classifiers associated with color keywords have strong predictive power. There is evidence of overfitting (e.g., the lack of yellow triangles in the data set); classifiers for objects of specific colors (such as banana) have more power than classifiers for general terms like "thin."

be classified—shows consistent improvement over manual labeling.

A. *Quality of Vision/Language Model*

The quality of the final grounded language model—the learned model of the relationship between language and percepts—is a product of the association between language tokens and the trained visual classifiers. Ideally, attribute descriptions should be associated primarily with a single classifier with good predictive power.

The evaluation was conducted on a corpus of 240 distinct RGB-D images of 50 objects, or *scenes*. 200 images were used to train color classifiers, and 40 images were used for testing. Results are reported on four colors because the other colors had too few instances to be divided into training and test data.

Figure 2 shows the accuracy of learned classifiers on objects described as being certain colors. Classifiers for color keywords are strongly predictive on objects that are those colors. Because of the simplicity of objects used, some overfitting occurs (*e.g.*, the classifier named "banana" has some predictive power on yellow objects). Classifiers corresponding to keywords which are not associated with colors (such as "half") have generally low confidence.

B. *Accuracy and Annotation Effort*

The evaluation was conducted on a corpus of 225 distinct RGB-D images of 45 objects, which we categorized into six rough color groupings. 125 images were used to train color classifiers as described in Section III-B, with a held-out set of 100 objects for testing. Figure 3 shows the accuracy of the trained classifiers on the test set as the size of the training set is increased. Results vary substantially by participant, reinforcing the belief that additional training by individual end users may be beneficial.

The largest source of errors in perception system was the variation in shades of colors per color contained in the training data—blue objects varied more in actual color than red, for example. The next largest class of errors is overfit to the training data based on the small set of objects. Perception system errors were largely between red/yellow and blue/-green objects; we attribute these to a combination of lighting conditions and the variation in shades of colors learned by classifiers.

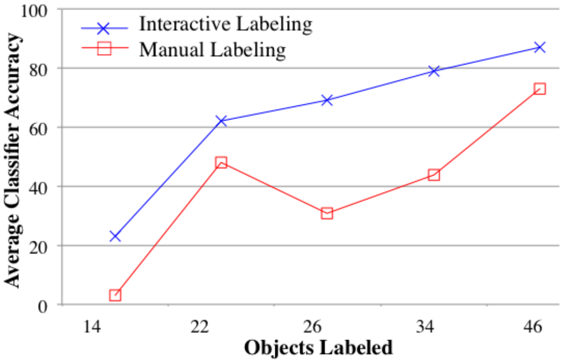

Fig. 3. Number of objects correctly classified in a held-out test set of 100 object/label pairs (y axis), per samples provided by a human annotator (x axis). After an initial set of 14 samples (to ensure the relevant classifiers exist), the active learning model consistently outperforms naive manual labeling.

V. User Study

The system was evaluated objectively for reduced labeling with an interactive system. We also conducted the pilot (described above in Section III-A) with a small number of annotators to learn about ease of use, acceptance, frustration and some general concepts with different kinds of labeling. [1]

A. *Manual Labeling Effort*

In our study, we trained classifiers and collected descriptions on 225 visual scenes (a scene being a single RGB-D view of one object). As described in Section III-A, the interactive learner attempted to classify objects in scenes as training progressed, asking for descriptions only when needed.

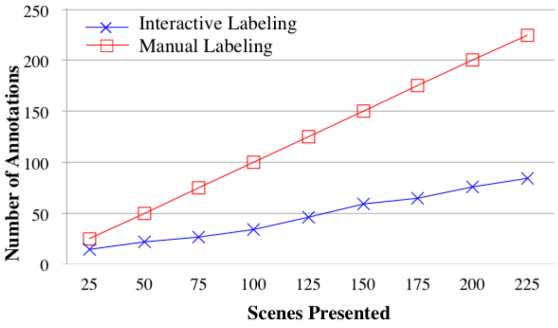

Fig. 4. The number of annotations (y axis) used to train classifiers on a certain number of visual scenes (x axis). Interactive labeling, which used classifier confidence to avoid requesting descriptions of some scenes, was able to request 63% fewer annotations than manual labeling. In the remaining cases, verification questions were asked. All comparative user results reflect these differences.

Overall, the system requested 84 descriptions, and correctly classified 141 objects in scenes without human labeling (see Figure 4), 37% of the annotator effort that would be required to label all 225 scenes.

B. *Comparative Study*

In this user experience study, we asked the participants to label ten items in each of two cases: manual labeling (they describe each object), and interactive labeling (they describe new or uncertain objects based on prompts from the system).They were then asked to provide their feedback on both approaches. The survey questionnaire was designed to measure comfort, efficiency, ease and accuracy of the systems. The participants rated their preference on a 5-point Likert-type scale, ranging from Strongly Agree to Strongly Disagree. A summary of the results can be found in Figure 5.

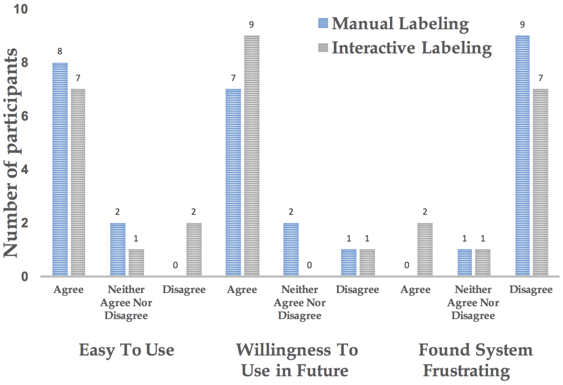

Fig. 5. User study comparison of ease of use, willingness to use the labeling system, and level of frustration associated with both the interactive labeling method and the manual labeling method.

Because of the increased number of questions asked, users found manual labeling slightly easier to use than interactive labeling. Users also favored interactive labeling when expressing their willingness to use the system in future if a robot needs to learn similar data.

Finally, frustration level with the system was queried to determine the state of mind of participants using each method. For interactive learning, only 10% (one user) was ‘dissatisfied’ and ‘very frustrated’. Anecdotally, this user was reacting to the system being “annoyingly polite,” e.g., wordy and apologetic in its interactions. Demographic difference may be a contributing factor in this case, as other users expressed no annoyance.

Another source of frustration was the lag time of the interactive labeling system; as classifiers were trained, users had to wait as much as 30 seconds for a response. User expectation of responsiveness is high, especially when the surrounding world has achieved remarkable progress in the field of robot interaction.

C. *General Questions*

Users were asked some general comparison questions to know their preference about the robot systems. Each user was asked to express their preferred

[1] See `http://tiny.cc/iral-al` for a brief video example of this interaction.

TABLE I
QUESTIONS ASKED COMPARING INTERACTIVE *vs.* MANUAL LABELING AND RESPONSES.

Questions	**Agree**	**Neither Agree Nor Disagree**	**Disagree**
Teaching was easier with interactive labeling as opposed to manual labeling.	6	1	3
Humans learn much faster and more accurately compared to interactive labeling.	6	3	1
Comfort level with Interactive labeling was higher than that with manual labeling.	4	3	3
An interactive system is much more fun than a monotonous one.	8	2	0

system to teach robot. Most of the participants preferred interactive labeling (see Table I).

VI. CONCLUSION AND FUTURE WORK

Ten objects cannot be considered sufficient for experiments to determine comfort and ease of use, and experimenting with more objects will help get a better idea on the willingness of users to teach a robot. Broadly speaking, however, these experiments support the intuition that interactive labeling will be more pleasant and more efficient, given careful interface design. The results also support the idea that actively seeking language labels can lower human labor costs in teaching robots.

In future, this work will extend to much large sample sizes, as well as incorporating active learning on a more 'real world' learning problem, in which the labels to be learned go beyond color to incorporate non-visual attributes (e.g., shape, weight), and into object recognition. We plan to incorporate active learning to establish spatial relationships between objects, as difficulty of manual labeling increases as the complexity of the symbolic representation increases.

Additionally, we will utilize the entropy of the classifier from the confidence estimate of the classifier as the classification deciding factor instead of user input. We will not employ verification by user if the classifier probability is too high for the object. We also believe that incorporating multimodal interaction, including speech recognition and gesture recognition [13]. We also conclude that the user expectations of speed and naturalness of interaction are really high in case of an interactive system.

This work is aimed at demonstrating that incorporating active learning in the collection of data annotation can reduce human labeling cost and learning time, and improve user comfort level. We show that a robot driven system can reduce manual labeling costs on a small problem, and efficiently learns from untrained users. While the presented system is in the early stages of development, initial results show that this is a promising area of research; as grounded language acquisition is a fundamental problem in human robot interactions, achieving the same level of learning with less effort is a crucial target.

This work is aimed at demonstrating that incorporating active learning in the collection of data annotation can reduce human labeling cost and learning time, and improve user comfort level. We show that a robot driven system can reduce manual labeling costs on a small problem, and efficiently learns from untrained users. While the presented system is in the early stages of development, initial results show that this is a promising area of research; as grounded language acquisition is a fundamental problem in human robot interactions, achieving the same level of learning with less effort is a crucial target.

REFERENCES

[1] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical Matching Pursuit for image classification: architecture and fast algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, December 2011.

[2] Maya Cakmak and Andrea L Thomaz. Designing robot learners that ask good questions. In *Proceedings of the 7^{th} annual ACM/IEEE international conference on Human-Robot Interaction*, pages 17–24. ACM, 2012.

[3] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. Designing interactions for robot

active learners. *IEEE Transactions on Autonomous Mental Development*, 2(2):108–118, 2010.

[4] Maya Cakmak, Nick DePalma, Rosa I Arriaga, and Andrea L Thomaz. Exploiting social partners in robot learning. *Autonomous Robots*, 29 (3-4):309–329, 2010.

[5] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 1996.

[6] Felix Duvallet, Matthew R Walter, Thomas Howard, Sachithra Hemachandra, Jean Oh, Seth Teller, Nicholas Roy, and Anthony Stentz. Inferring maps and behaviors from natural language instructions. In *Experimental Robotics*, pages 373–388. Springer, 2016.

[7] Goren Gordon and Cynthia Breazeal. Bayesian active learning-based robot tutor for children's word-reading skills. In *AAAI*, pages 1343–1349, 2015.

[8] Sachithra Hemachandra and Matthew R Walter. Information-theoretic dialog to improve spatial-semantic representations. In *Intelligent Robots and Systems (IROS)*. IEEE, 2015.

[9] W Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a robot via human feedback: A case study. In *Social Robotics*, pages 460–470. Springer, 2013.

[10] Johannes Kulick, Marc Toussaint, Tobias Lang, and Manuel Lopes. Active learning for teaching a robot grounded relational symbols. In *IJCAI*, 2013.

[11] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297, 1967.

[12] Cynthia Matuszek*, Nicholas FitzGerald*, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *29th International Conference on Machine Learning (ICML)*, Edinburgh, Scotland, June 2013.

[13] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proc. of the 28th National Conference on Artificial Intelligence (AAAI)*, March 2014.

[14] Raymond J. Mooney. Learning to connect language and perception. In Dieter Fox and Carla P. Gomes, editors, *Proc. of the 23th AAAI Conf. on Artificial Intelligence, AAAI 2008*, pages 1598–1601, Chicago, Illinois, July 2008. AAAI Press.

[15] Morgan Quigley, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibsz, Eric Berger, Rob Wheeler, and Andrew Ng. ROS: An Open-source Robot Operating System. In *Open-source Software Workshop of the International Conference on Robotics and Automation*. IEEE, 2009.

[16] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text Mining*, pages 1–20, 2010.

[17] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1–4. IEEE, 2011.

[18] Danijel Skočaj, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M Kruijff, Marc Hanheide, Nick Hawes, Jeremy L Wyatt, Thomas Keller, Kai Zhou, et al. An integrated system for interactive continuous learning of categorical knowledge. *Journal of Experimental & Theoretical Artificial Intelligence*, 2016.

[19] Stefanie Tellex and Deb Roy. Towards surveillance video search by natural language query. In *Conference on Image and Video Retrieval (CIVIR-2009)*, 2009.

[20] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth Teller, and Nicholas Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI magazine*, 32 (4):64–76, 2011.

[21] Stefanie Tellex, Pratiksha Thaker, Robin Deits, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. Toward information theoretic human-robot dialog. *Robotics*, 2013.