Continuously Improving Natural Language Understanding for Robotic Systems through Semantic Parsing, Dialog, and Multi-modal Perception

Jesse Thomason
The University of Texas at Austin
jesse@cs.utexas.edu

Doctoral Dissertation Proposal

November 23, 2016

Abstract

Robotic systems that interact with untrained human users must be able to understand and respond to natural language commands and questions. If a person requests "take me to Alice's office", the system and person must know that Alice is a person who owns some unique office. Similarly, if a person requests "bring me the heavy, green mug", the system and person must both know "heavy", "green", and "mug" are properties that describe an object in the environment, and have similar ideas about to what objects those properties apply. To facilitate deployment, methods to achieve these goals should require little initial in-domain data.

We present completed work on understanding human language commands using sparse initial resources for semantic parsing. Clarification dialog with humans simultaneously resolves misunderstandings and generates more training data for better downstream parser performance. We introduce multi-modal grounding classifiers to give the robotic system perceptual contexts to understand object properties like "green" and "heavy". Additionally, we introduce and explore the task of word sense synonym set induction, which aims to discover polysemy and synonymy, which is helpful in the presence of sparse data and ambiguous properties such as "light" (light-colored versus lightweight).

We propose to combine these orthogonal components into an integrated robotic system that understands human commands involving both static domain knowledge (such as who owns what office) and perceptual grounding (such as object retrieval). Additionally, we propose to strengthen the perceptual grounding component by performing word sense synonym set induction on object property words. We offer several long-term proposals to improve such an integrated system: exploring novel objects using only the context-necessary set of behaviors, a more natural learning paradigm for perception, and leveraging linguistic accommodation to improve parsing.

Contents

1 Introduction

As robots become more pervasive in human environments, they must be able to understand questions and commands from untrained human users in natural language. Consider, for example,

Go to Alice's office and get the light mug for the chair. (1)

Human utterances like the one above can be translated into semantic meanings. Given a semantic meaning, a robot can check against its knowledge and perception to resolve references to the real world and take actions appropriately in response. For example, one semantic interpretation of the above is

$$\text{go}(\text{the}(\lambda x.(\text{office}(x) \wedge \text{owns}(\text{alice}, x)))) \wedge \text{deliver}(\text{the}(\lambda y.(\text{light}_2(y) \wedge \text{mug}_1(y))), \text{bob}) \quad (2)$$

Translating human utterances to semantic meanings helps overcome synonymy of commands and words, compositionality (e.g. "Alice's office", "not green"), and ambiguity (e.g. "the chair" for furniture and "the chair" for the head of an organization). For example, in (1) above, "the chair" refers to a person, `bob`, not an object, and "Alice's office" is understood as a request for some space satisfying both being an `office` and belonging to `alice`.

In order to converse about the environment they share with humans, these robots must gather and maintain world knowledge through perception. Some world knowledge is ontological, such as the layout of a building, ownership relations between people and rooms, or assignments between patients and doctors in a hospital. This information can be created by humans and stored as a knowledge base accessible for language understanding. For example, in (1), the parse of "Alice's office" can be grounded against such a knowledge base to find the room satisfying the given constraints. Other world knowledge is perceptual, such as whether an object is a "mug", where some movable objects were last seen, and whether an object can be picked up and moved somewhere else. A service robot in a human environment needs both types of knowledge to understand and respond to human requests through dialog and actions.

The words used to describe object properties do not form a one-to-one mapping with underlying predicates. For example, "claret" and "purple" can reasonably refer to the same underlying visual classifier. Additionally, "light" may refer to either a predicate for light coloration or a predicate for light weight. Robust robot perception must account for these ambiguous word senses, and should benefit from identifying synonymous senses. In (1), for example, the system must find the correct sense of "light", marked as a second sense in the associated parse (2).

In this proposal, we describe steps towards a system that improves its language understanding and perception components incrementally through dialog with a human user. Such a system would have a natural language understanding pipeline comprising semantic parsing, dialog, and perception. For learning, we discuss: using existing resources like ImageNet (Deng et al. [2009]) together with small, hand-annotated semantic information (lexicon, ontology) to initialize a system; leveraging interactive human-computer games like "I Spy" to further bootstrap robot perception; and continuous improvement of parsing and perception components once deployed through passive supervision from human dialog. Such a system would be able to understand and correctly execute the request in (1) while refining both its parsing and perception components through interaction with humans over time.

In the remainder of this document, we first discuss the substantial body of related work on semantic parsing for understanding human language commands, robot perception, and handling word senses (Section 2). We continue with a discussion of completed work that uses weak supervision from human-robot dialog to generate training data for a semantic parser (Section 3). We then describe completed work that grounds predicates in multi-modal perception, allowing a robot to move beyond pre-written predicates ("office", "possesses") into human descriptor words ("mug", "heavy") (Section 4). We continue in this vein, discussing a methodology for synonym set induction (Section 5), which we propose to apply to perceptual predicates (Section 6.1). We propose to integrate all these components into a single robotic system for understanding and grounding human language requests (Section 6.2). Finally, we discuss longer-term proposals for improving an integrated system (Section 7) and summarize this proposal (Section 8).

2 Background and Related Work

This proposal concerns the integration of semantic parser learning with robot perception for natural language understanding without requiring the use of large corpora or initial domain-relevant training data. Completed work as well as that proposed is situated within the Building-Wide Intelligence (BWI) project at the University of Texas at Austin[1]. We use Segway-based robots for embodied experiments as described in Khandelwal et al. [2014].

We discuss existing work on instructing robots through natural language, a task our proposed integrated system will do with semantic parsing and perception. We overview relevant work on learning semantic parsers, including work on inducing training data for semantic parses from conversations. We discuss language grounding as a task, grounding in machine perception, and grounding with additional signal from human-robot interactions. Finally, we overview natural language understanding tasks involving detecting and handling polysemy and synonymy in language, which is relevant for our integration goal of using synset induction to improve perceptual grounding performance.

2.1 Instructing Robots in Natural Language

Instructing robots through natural language is essential for the cooperation of humans and robots in shared environments. Researchers have focused on different aspects of human-robot communication, including using perception alongside semantic parsing for action understanding and acquiring new actions from language descriptions in a perceivable environment.

Understanding the mutual environment is essential for human-robot communication. Semantic parsing has been used as the understanding step in tasks like unconstrained natural language instruction where a robot must navigate an unknown environment (Kollar et al. [2010], Matuszek et al. [2012b]). Weak supervision can be used to improve these parsers continuously based on interactions with humans (Artzi and Zettlemoyer [2013b]), similar to the goals of this proposal. Simpler parsing approaches, such as transforming commands using semantic role labeling to form

[1] `http://www.cs.utexas.edu/~larg/bwi_web/`

a meaning representation, are less training intensive at the cost of being less robust to language variation (Bastianelli et al. [2013]). There have been focused efforts to relate to understand human language commands with respect to a shared environment, such as the SemEval task of Dukes [2014]. Work on semantic graphs connects environment referents probabilisticly based on both sensor data and human language (Walter et al. [2013]), while similar work additionally incorporates knowledge base information and conversation context (Mohan et al. [2013]). One framework acts to both understand human language requests about objects in the world and generate language requests regarding the shared environment (Tellex et al. [2014]). Other work performs perception first to build a model of the shared environment, then performs semantic parsing independently and uses the perceived world as a knowledge base to resolve predicate information (Yang et al. [2014], Lu and Chen [2015]).

Recent work angles to translate human instructions directly to grounded behavior like route-following, skipping parsing in favor of sequence-to-sequence, instruction-to-action mapping using neural methods (Mei et al. [2016]). Past methods consider information jointly from the instructional utterance and the perceived environment to perform action understanding as a sequence (Misra et al. [2014]) or hierarchy (Kuehne et al. [2014]).

Going beyond action understanding, past work has also used semantic representations of utterances together with perception of objects in an environment to learn new manipulation behaviors from human instruction (She et al. [2014]). Similarly focused situated action learning for navigation maps human language instructions into executable program-like behaviors that can be used as modules for hierarchically-composed actions (Meriçli et al. [2014]).

We propose a robotic system that understands requests for actions in natural language that can include both domain knowledge and perceptual information. This will involve semantic parsing as part of understanding, but does not include plans for action learning. Instead, we focus on executing pre-programmed actions (such as delivery and navigation) robust to language variations and use of perceptual predicates like "heavy mug" to describe objects in the real world.

2.2 Learning Semantic Parsers

A semantic parser, for example Artzi and Zettlemoyer [2013a], is a function from strings of words to a semantic meaning defined by some ontology. Formally, a parser $P : \mathbb{P}(W) \times L_O \rightarrow S_0$ takes in a sequence of word tokens $T \in \mathbb{P}(W)$ for W the set of all word tokens and a lexicon L_O for ontology O and outputs a semantic parse $s \in S_O$ the set of all semantic parses in ontology O. An ontology O defines a set of atoms and predicates. Atoms are things like items, places, people, and true/false boolean values. Predicates are functions on atoms that return other atoms (such as true and false values). The lexicon L is a data structure that contains information about how individual word tokens relate to that ontology, for example that token "alice" refers to ontological atom `alice` or that possessive marker "'s" invokes predicate `owns` (see Figure 1). A semantic parse is a meaning representation in terms of ontological predicates and lambda expressions. The meaning of "bring alice a coffee", for example, could be represented as

$$\text{bring}(\text{alice}, \text{coffee})$$

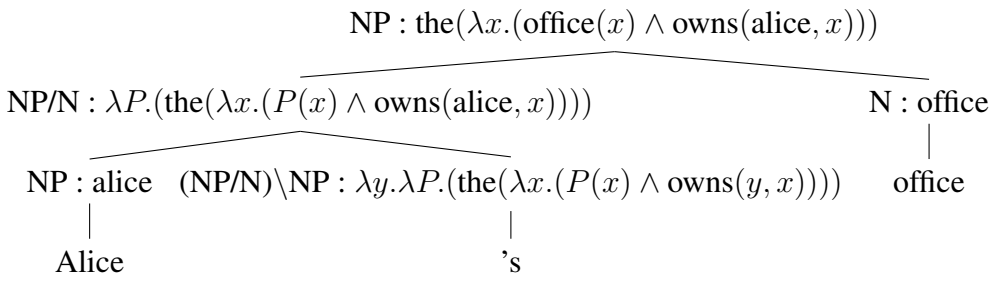


Figure 1: A CCG-driven λ-calculus parse of the expression "Alice's office".

A lambda expression allows meaning to abstract over sets of ontological atoms, predicates, and truth values. The meaning of "go to alice's office" could be represented as

$$\text{go}(\lambda x.(\text{the}(\text{office}(x) \wedge \text{owns}(\text{alice}, x))))$$

In this case, the predicate *the* will pick out some unique atom x that is both an `office` and belongs to `alice`, returning that atom as the argument of the `go` command.

Learning a semantic parser to perform these translations is non-trivial, but using statistical parsing with compositional semantics makes it feasible. Many works, including our own, frame compositional semantics through combinatory categorial grammar (CCG) (Steedman and Baldridge [2011]). CCG adds a syntactic category to each token of input sequence t and then hierarchically combines those categories to form a syntax tree. These syntax categories are typically a small set of basic, like N (noun) and NP (noun phrase), together with compositional entries, like PP/NP (a prepositional phrase formed after consuming a noun to the right). In a compositional semantics framework, the input lexicon L contains entries mapping word tokens to not just semantic meanings but also CCG syntax categories. Semantic meanings can be composed according to the hierarchy established by the CCG syntactic parse to form a semantic parse that spans the entire token sequence. Figure 1 demonstrates this for "Alice's office".

Given data in the form of paired token sequences with their semantic parse trees, a statistical parser can be trained to produce those parses given novel sequences (Berant et al. [2013], Liang and Potts [2015]). However, annotating token sequences with entire parse trees is costly, and many works instead train with latent trees, requiring only the final desired semantic form (tree root) (Liang et al. [2011]). Some works go beyond parser training and additionally incorporate ontology matching, which removes the restriction that the lexicon use predicates that exactly match the domain (Kwiatkowski et al. [2013]).

The number of parameters to be learned in a semantic parser becomes large quickly even in a restricted domain, and annotating natural language with semantic forms is an expensive human effort. Consequently, some work in semantic parser learning has focused on overcoming data sparsity during training, such as inducing training examples automatically from existing conversational data (Artzi and Zettlemoyer [2011]).

In our completed work (Section 3), we build on that idea to gather parsing training examples from dialog. We propose to use semantic parsing as an understanding step in a robotic system capable of grounding language referents in world knowledge as well as perceptual information. Adding

perceptual information means we will have confidence from both the parser about a semantic form and also from the perceptual system in how likely that parse is given the objects in the real world. In Section 6.2, we propose novel contributions regarding this potential feedback loop.

2.3 Language Grounding with Machine Perception

Commanding robots with language requires both semantic understanding and a subsequent grounding step where referents in the real world are connected to language used to describe them. Mapping from referring expressions such as "the blue cup" to an object referent in the world is an example of the *symbol grounding problem* (Harnad [1990]). Symbol grounding involves connecting internal representations of information in a machine to real world data from its sensory perception. *Grounded language learning* bridges these symbols with natural language. Comparative studies have established that joint representations of language that consider some form of perception outperform text-only representations of word meaning (Silberer and Lapata [2012]).

Early work used vision together with speech descriptions of objects to learn grounded semantics (Roy and Pentland [2002]). Recently, most work has focused on combining language with visual information. For grounding referring expressions in an environment, many learn perceptual classifiers for words given some pairing of human descriptions and labeled scenes (A. Lazaridou and Baroni. [2014], Sun et al. [2013]). Some approaches additionally incorporate language models into the learning phase (Krishnamurthy and Kollar [2013], FitzGerald et al. [2013], Matuszek et al. [2012a]). Some researchers have translated images into a distribution over possible descriptions, attempting to solve the problem in the other direction first, then doing query similarity in that textual space (Guadarrama et al. [2015]).

Recent work bypasses any explicit language understanding in favor of neural methods, such as Hu et al. [2016], who localize an object in a given image given a target query in natural language. In a related task, other recent work aims to resolve ambiguities like prepositional phrase attachment in natural language by using associated images to gather additional information (Christie et al. [2016]).

There has been some work on combining language with sensory modalities other than vision, such as audio (Kiela and Clark [2015]). Additionally, researchers have explored the use of haptic and proprioceptive feedback from a robot arm to automatically learn to order objects by weight, height, an width (Sinapov et al. [2016]). Other works bypass perception and work with knowledge base structures directly, learning to map streams of text references of world states knowledge base entries describing those states (Liang et al. [2009]).

In our completed work (Section 4), we introduce multi-modal perception for a robotic system using vision together with haptics, audio, and proprioception. We propose to integrate that perception with an embodied system that accepts and understands natural language commands. Thus, we will use multi-modal perception to ground language predicates like "heavy" used by humans in descriptions of objects in the environment.

2.4 Language Grounding with Human-Robot Interaction

Machine perception is not only necessary for human-robot interaction; it can also be improved by that interaction. A number of researchers have focused on solving the symbol grounding problem for situated robots by leveraging their interactions with the very humans they are working to understand.

One line of existing work focuses on gathering data from human demonstration and speech to learn language grounding. These use an existing corpus of human demonstration as input. One work uses unscripted human descriptions of objects together with their deictic hand gestures to train a grounding system for identifying referent objects (Matuszek et al. [2014]). Similar work used only speech from demonstrations from humans describing objects to achieve one-shot learning of object attributes and names (Perera and Allen [2013]). Other researchers have focused on learning unary properties of objects ("red") together with relational ("taller") and differentiating ("differ by weight") properties of objects by exploring them with a robotic arm provided properties and relational labels as human supervision after that exploration (Sinapov et al. [2014b]).

Closer to the work in this proposal, some researchers gather data for perceptual grounding using interaction with a human interlocutor. This combination of dialog and perception affords new opportunities for smart interactions, such as the robot asking questions targeting weaknesses in its understanding (as in our Thomason et al. [2016]). Early work on learning to ground object attributes and names using dialog framed the data gathering phase as a "20 Questions"-style game (Vogel et al. [2010]) where a robot tried to guess a target object by asking narrowing questions (e.g. "is it red?"). Contemporary to this, other research focused on acquiring the same attribute and name perceptual understandings through a command-, rather than game-based environment (Dindo and Zambuto [2010]). Researchers have carried this idea to more complete systems with both perceptual grounding and action learning capability for identifying and manipulating objects, where the agent can request more information about uncertain concepts (Mohan et al. [2012]). Similar to other work that learns from demonstration and description offline, Kollar et al. [2013] studies the joint acquisition of perceptual classifiers and language understanding in an interactive setting. Focused efforts have begun studying one-shot object attribute learning (Krause et al. [2014]).

More recent work aims to address perceptual mismatch between humans and robots, since our sensory systems differ, and delineations present in the one may be undetectable in the other (Liu et al. [2014], Liu and Chai [2015]). In the vein of game-based data gathering, researchers have framed learning attribute classifiers for objects as an "I Spy" game in which a human describes a target object among several options to a robot and confirms when the correct one is identified (Parde et al. [2015]). Other object identification work has focused on integrating language with gesture, bypassing perception in favor of language co-occurrences with particular objects (Whitney et al. [2016]).

In completed work, we bootstrap our perception system using an interactive "I Spy" game (Section 4). We propose to introduce continuous learning to the perceptual component of a robotic system. As the robot has dialog interactions with humans involving requests like "bring bob the heavy mug", once the correct object has been identified and delivered, that object can serve as a positive example for "heavy" and "mug" perceptual classifiers. This is similar to the existing feedback loop in our completed work, where dialog confirmations allow us to generate additional

semantic parser training data from earlier conversational misunderstandings (Section 3).

2.5 Polysemy and Synonymy in Language Understanding

Semantic understanding in language is complicated by words that have multiple, distinct meanings, and by sets of words with the same underlying meaning. A sense inventory for words, such as WordNet (Fellbaum [1998]), structures word meaning into senses which can be taken on by one or more words. Words that refer to the same sense are called synonymous. A word which refers to multiple meanings across its different senses is polysemous. Sets of word senses are called *synsets*.

Word sense induction (WSI) is the task of determining whether an individual word type, such as "bat", is polysemous, and what its underlying senses are. Given a set of word senses and a word in context, the task of *word sense disambiguation* (WSD) is to choose what sense the word expresses in the given context (Navigli [2009]). WSD has applications across language understanding, since the multiple meanings of a word can be arbitrarily different and unrelated.

Traditional work on WSI discovers senses for a word by clustering the textual contexts in which it occurs (Yarowsky [1995], Pedersen and Bruce [1997], Schutze [1998], Bordag [2006], Navigli [2009], Manandhar et al. [2010], Reisinger and Mooney [2010], Di Marco and Navigli [2013]). The multiple meanings for "bat" can be recognized as two clusters: one a set of contexts with words like "cave" and "flew"; and another with words like "baseball" and "strike". Other notions of context can be used to discover word senses, such as images the word is used to describe. Some previous work has recognized the value of perceptual word senses for tasks such as image-segment labeling (Barnard and Johnson [2005]) and image retrieval from text queries (Lucchi and Weston [2012]).

Past work has used visual information to disambiguate word senses, but assumes the senses of each word are known in advance (Barnard and Johnson [2005]). Using both textual and visual information to perform WSI has been done, but on datasets where every input word is known in advance to be polysemous (Loeff et al. [2006], Saenko and Darrell [2008]).

Recent work performs co-clustering in separate textual and visual spaces, treating textual clusters as word senses and visual clusters as iconographic senses (viewpoint changes, color differences, etc.) that offer a finer-grained distinction than word senses (Chen et al. [2015]).

We propose to use polysemy and synonymy detection to induce synsets for predicates people use to describe objects to our robotic system. This should allow us to discover, for example, the polysemous meanings of "light" (weight and color). Additionally, it will create stronger classifiers for rare words like "claret" by combining them with more common words like "purple", mapping these surface forms to the same underlying synset-based classifier.

3 Learning to Interpret Natural Language Commands through Human-Robot Dialog

Intelligent robots need to understand requests from naive users through natural language. Here we discuss a dialog agent for mobile robots that understands human instructions through seman-

TASK TO COMPLETE

Dave Daniel wants the item in slot 5.

ROBOT	How can I help?
YOU	please bring the item in slot 5 to dave daniel
ROBOT	To whom should I bring something?
YOU	dave daniel
ROBOT	What action did you want me to take involving dave daniel?
YOU	5
ROBOT	What should I bring to dave daniel?
YOU	a calander
ROBOT	I'm sorry, but I couldn't pinpoint what you meant by that.
ROBOT	What should I bring to dave daniel?
YOU	a day planner
ROBOT	I'm sorry, but I couldn't pinpoint what you meant by that.
ROBOT	What should I bring to dave daniel?
YOU	calendar
ROBOT	You want me to bring calendar to dave daniel?
YOU	yes
ROBOT	I thought so
ROBOT	Happy to help
YOU	

Final Task

DIRECTORY

People:

Alice Ashcraft; Secretary
Francis ("Frannie") Foster
Robert ("Bob") Brown
Carol Clark, PhD
Dave Daniel, PhD
George Green; Intern
Evelyn ("Eve") Eckhart
Mallory Morgan; Director
Peggy Parker, PhD
Walter Ward; Supervisor

Items available to robot:

Figure 2: This abridged conversation took place when the system had access to only the initial, small data (no additional training examples yet).

tic parsing, actively resolves ambiguities using a dialog manager, and incrementally learns from human-robot conversations by inducing training data from user paraphrases. Figure 2 gives an example conversation when the dialog agent had only slim initial training data for language understanding. The clarification questions in this conversation let the agent learn that "calender" and "planner" mean "calendar". This completed work allows us to translate utterances into logical forms, and is the first step to resolving (1) in our proposed work to integrate parsing and perception in an embodied robotic system. Full details are available in Thomason et al. [2015].

3.1 Methods

A human user first gives a command to our dialog agent, then the agent can ask clarification questions (Figure 3). The agent maintains a belief state about the user's goal. When it is confident in this state, the dialog ends and the goal is passed on to the robot or other underlying system.

The agent produces a semantic form for each user utterance. We use the University of Washington Semantic Parsing Framework (SPF) (Artzi and Zettlemoyer [2013a]), a state-of-the-art system for mapping natural language to meaning representations using λ-calculus and combinatory categorial grammar (CCG).

To get the system "off the ground" we initialize the parser with a small seed lexicon and then train it on a small set of supervised utterance/logical-form pairs. We use a seed lexicon of 105 entries (40 of which are named entities) and a training set of only 5 pairs.

The agent maintains a belief state about the user goal with three components: `action`, `patient`, and `recipient`. Each component is a histogram of confidences over possible assignments. The agent supports two actions: walking and bringing items, so the belief state for `action` is two confidence values in $[0, 1]$. `recipient` and `patient` can take values over the space of entities (people, rooms, items) in the knowledge base as well as a null value $\varnothing$.

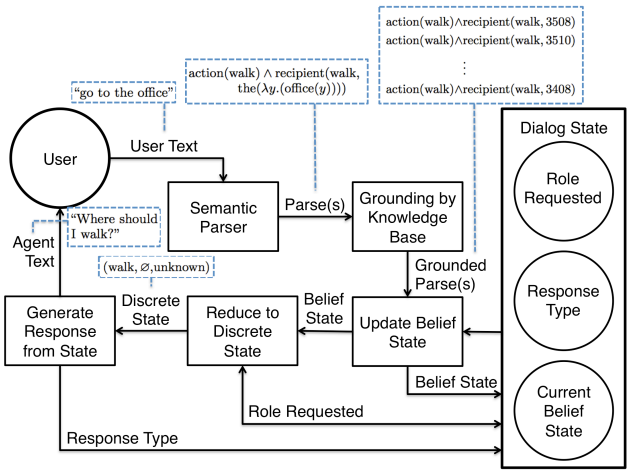


Figure 3: Dialog agent workflow processing user command "go to the office".

Multiple meaning hypotheses may be generated from a user utterance. Consider:

expression go to the office

logical form $\text{action}(\text{walk}) \wedge \text{recipient}(\text{walk}, \text{the}(\lambda y.(\text{office}(y))))$

For n offices, this logical form has n groundings producing different meanings (see Figure 3). The agent can be confident that walking is the task, but its confidence in the n meanings for `recipient` is weakened. We use a confidence update based on the number k of hypotheses generated to track the agent's confidence in its understanding of each component of the request. For a user-initiative (open-ended) statement like this one, the agent updates all components of the belief state. For each candidate hypothesis $H_{i,c}$, with $0 \leq i < k$, $c \in \{\texttt{action}, \texttt{patient}, \texttt{recipient}\}$, the agent updates:

$$\text{conf}(c = H_{i,c}) \leftarrow \text{conf}(c = H_{i,c}) \left(1 - \frac{\alpha}{k}\right) + \frac{\alpha}{k}$$

Where $0 < \alpha < 1$ is the threshold of confidence above which the candidate is accepted without further clarification. The confidence in unmentioned arguments is decayed to wash out previous misunderstandings. For A_c, the set of all candidates of component c, $\bar{A}_c = A_c \setminus \cup_i \{H_{i,c}\}$ are unmentioned. For each $\bar{H}_{j,c} \in \bar{A}_c$, the agent updates:

$$\text{conf}(c = \bar{H}_{j,c}) \leftarrow \gamma \text{conf}(c = \bar{H}_{j,c})$$

where $0 \leq \gamma \leq 1$ is a decay parameter.

System-initiative responses are associated with a particular requested component. These can take the form of confirmations or prompts for components. For the former, user affirmation will update the confidence of all mentioned values to 1. For the latter, the positive and negative updates described above operate only on the requested component.

The agent uses a static dialog policy π operating over a discrete set of states composed of `action`, `patient`, `recipient` tuples together with the role to be clarified. The agent's continuous belief state S is reduced to a discrete state S' by considering the top candidate arguments T_c for each component c:

$$T_c = \text{argmax}_{t \in A_c}(\text{conf}(c = t))$$

Table 1: Representative subset of our policy π for mapping discrete states S' to questions.

S'		$\pi(S')$	
(`action`, `patient`, `recipient`)	**Role Request**	**Response**	**Initiative**
(unknown unknown, unknown)	all	Sorry I couldn't understand that. Could you reword your original request?	user
(unknown, $T_{\texttt{patient}}$, $T_{\texttt{recipient}}$)	action	What action did you want me to take involving $T_{\texttt{patient}}$ and $T_{\texttt{recipient}}$?	system
(walk, $\varnothing$, unknown)	recipient	Where should I walk?	system
(bring, unknown, $T_{\texttt{recipient}}$)	patient	What should I bring to $T_{\texttt{recipient}}$?	system
(walk, $\varnothing$, $T_{\texttt{recipient}}$)	confirmation	You want me to walk to $T_{\texttt{recipient}}$?	system
(bring, $T_{\texttt{patient}}$, $T_{\texttt{recipient}}$)	confirmation	You want me to bring $T_{\texttt{patient}}$ to $T_{\texttt{recipient}}$?	system

Each component c of S' is selected by choosing either T_c or "unknown" with probability conf($\texttt{c} = T_c$). The component c with the minimum confidence is chosen as the role to request. If "unknown" is chosen for every component, the role requested is "all". If "unknown" is chosen for no component, the role requested is "confirmation". Some policy responses are given in Table 1. If each of the confidence values inspected during this process exceeds α, the conversation concludes. In all experiments, parameters $\alpha = 0.95, \gamma = 0.5$ were used.

Our agent induces parsing training examples from conversations with users to learn new lexical items. It uses dialog conclusions and explicit confirmations from users as supervision. The semantic parser in Figure 2 does not know the misspelling "calender", the word "planner", or number "5". When the user requests "item in slot 5" be delivered, it only confidently detects the `action`, "bring", of the user's goal. The `recipient`, "Dave Daniel", is clarified by a system-initiative question. When the agent asks for confirmation of the `action`, the user does not deny it, increasing the agent's confidence. While clarifying the `patient`, the user implicitly provides evidence that "calender", "planner", and "calendar" are the same. When two or more phrases are used in the same sub-dialog to clarify an argument, the eventual logical form selected is paired with the earlier surface forms for retraining.

User-initiative responses generate similar alignments. One users' conversation began "please report to room 3418", which the agent could not parse because of the new word "report". The agent understood the re-worded request "go to room 3418", and the former sentence was paired with the

logical form of this latter for training. When the retraining procedure explored possible semantic meanings for "report", it found a valid parse with the meaning of "go", "S/PP : $\lambda P.(\text{action}(\text{walk}) \wedge P(\text{walk}))$", and added it to the parser's lexicon. This meaning says that "report" should be followed by a prepositional phrase specifying a target for the walking `action`.

3.2 Experiments

We evaluated the learning agent in two contexts. We used Mechanical Turk to gather data from many diverse users asked to give the agent goals for an office environment. These users interacted with the agent through a web browser, but user expectations, frustrations, and lexical choices with a web browser versus a physical robot will likely differ. Thus, we also implemented an interface for the agent on a Segway-based robot platform (Segbot) operating on a floor of our university's computer science building.

We split the possible task goals into train and test sets. In both contexts, users performed a *navigation* and a *delivery* task. For the 10 possible navigation goals (10 rooms), we randomly selected 2 for testing. For the 50 possible delivery goals (10 people $\times$ 5 items), we randomly selected 10 for testing (80%/20% train/test split). The test goals for Mechanical Turk and the Segbot were the same, except in the former we anonymized the names of the people on our building's floor.

We ended all user sessions with a survey: "The tasks were easy to understand" (*Tasks Easy*); "The robot understood me" (*Understood*); and "The robot frustrated me" (*Frustrated*). For the Segbot experiment, we also prompted "I would use the robot to find a place unfamiliar to me in the building" (*Use Navigation*) and "I would use the robot to get items for myself or others" (*Use Delivery*). Users answered on a 5-point Likert scale: "Strongly Disagree"(0), "Somewhat Disagree"(1), "Neutral"(2), "Somewhat Agree"(3), "Strongly Agree"(4). Users could also provide open comments.

Mechanical Turk Experiments. The web interface shown in Figure 2 was used to test the agent with many users through Mechanical Turk. We performed incremental learning in batches to facilitate simultaneous user access. We assigned roughly half of users to the test condition and the other half to the train condition per batch. After gathering train and test results from a batch, we retrained the parser using the train conversation data. We repeated this for 3 batches of users, then we gathered results from a final testing batch in which there was no need to gather more training data. We used user conversations for retraining only when they achieved correct goals.

Navigation: Users were asked to send the robot to a random room from the appropriate train or test goals with the prompt "[person] needs the robot. Send it to the office where [s]he works". The referring expression for each person was chosen from: full names, first names, nicknames, and titles. In this task, the corresponding office number was listed next to each name, and the "items available" were not shown.

Delivery: Users were asked to tell the robot to assist a person with the prompt "[person] wants the item in slot [number]". The $(person, item)$ pairs were selected at random from the appropriate train or test goals. To avoid linguistic priming, the items were given pictorially (Figure 2).

For each train/test condition, we gathered responses from an average of 48 users per batch.

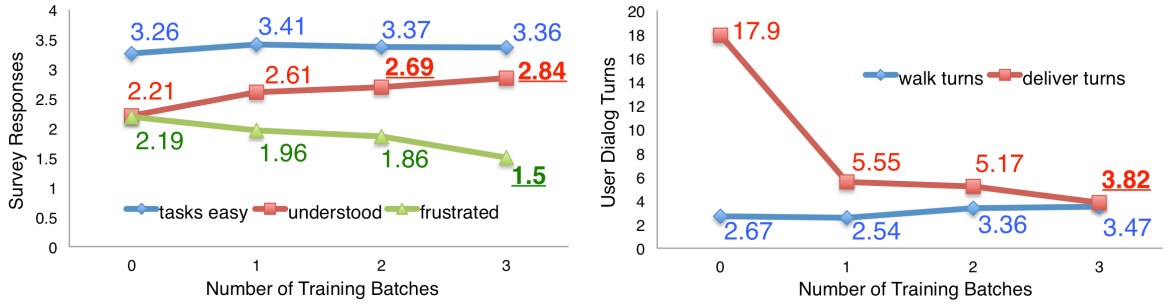


Figure 4: **Left**: Average Mechanical Turk survey responses across the four test batches. **Right**: Mean user turns in Mechanical Turk dialogs where the correct goal was reached. Means in underlined bold differ significantly ($p < 0.05$) from the batch 0 mean.

Figure 4 (Left) shows the mean survey-question responses across test batches. We used an unpaired Welch's two-tailed t-test to determine whether these means differed significantly. By batch 2, users felt that the agent understood them more than in batch 0. By batch 3, they felt that it frustrated them less. The dialog agent became more understandable and likable as a result of the semantic parser's learning, even though it had never seen the test-batch users' goals.

To determine whether learning reduced the number of utterances (turns) a user had to provide for the system to understand their goal, we counted user turns for dialogs where the user and agent agreed on the correct goal (Figure 4 (Right)). Learning successfully reduced the turns needed to understand multi-argument delivery goals.

With respect to users' free-form feedback, in testing batch 0, several enjoyed their conversations ("This was fun!! Wish it were longer!"). Several also commented on the small initial lexicon ("It was fun to try and learn how to talk to the robot in a way it would understand"). The responses by testing batch 3 had similarly excited-sounding users ("I had so much fun doing this hit!"). At least one user commented on the lexical variation they observed ("The robot fixed my grammatical error when I misspelled 'calender' Which was neat"). In addition to learning misspelling corrections and new referring expressions, the agent learned to parse things like "item in slot n" by matching n to the corresponding item and collapsing the whole phrase to this meaning.

Segbot Experiments. The agent was integrated into a Segway-based robot platform (Segbot) as shown in Figure 5 (Left) using the Robot Operating System (ROS) (Quigley et al. [2009]). The robot architecture is shown in Figure 5 (Right). Users interacted with the agent through a graphical user interface by typing in natural language. The agent generated queries to a symbolic planner formalized using action language $\mathcal{BC}$ (Lee et al. [2013]) from user goals.

For testing, users were given one goal from the navigation and delivery tasks, then filled out the survey. The task prompts included the directory panels used in the Mechanical Turk experiments pairing names and office numbers and showing items available to the robot for delivery (Figure 2).

We evaluated our agent's initial performance by giving 10 users one of each of these goals (so each delivery test goal was seen once and each navigation test goal was seen 5 times). Users were allowed to skip goals they felt they could not convey. We refer to this group as `Init Test`.

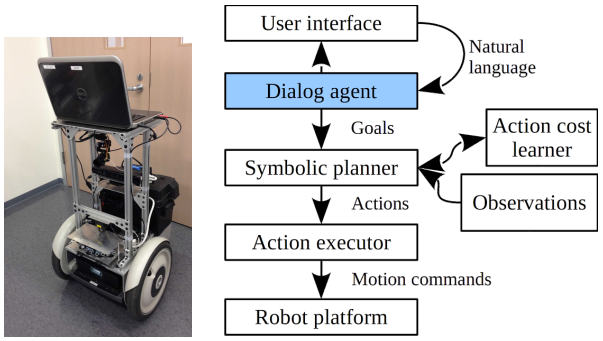


Figure 5: **Left:** Robot platform (Segbot) used in experiments. **Right:** Segbot architecture, implemented using Robot Operating System (ROS).

We then allowed the agent to perform incremental learning for four days in our office space. Students working here were encouraged to chat with it, but were not instructed on how to do so beyond a panel displaying the directory information and a brief prompt saying the robot could only perform "navigation and delivery tasks". Users in test conditions did not interact with the robot during training. After understanding and carrying out a goal, the robot prompted the user for whether the actions taken were correct. If they answered "yes" and the goal was not in the test set, the agent retrained its semantic parser with new training examples aligned from the conversation. [2]

We evaluated the retrained agent as before. The same testing goal pairs were used with 10 new users. We refer to this latter set as `Trained Test`.

During training, the robot understood and carried out 35 goals, learning incrementally from these conversations. Table 2 compares the survey responses of users and the number of goals users completed of each task type in the `Init Test` and `Trained Test` groups. We use the proportion of users having completed goals in each task as a metric for dialog efficiency. For navigation goals, `Init Test` had an average dialog length of 3.89, slightly longer than the 3.33 for `Train Test`.

We note that there is significant improvement in user perception of the robot's understanding, and trends towards less user frustration and higher delivery-goal correctness. Though users did not significantly favor using the robot for tasks after training, several users in both groups commented that they would not use guidance only because the Segbot moved too slowly.

4 Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy"

Grounded language learning bridges words like 'red' and 'square' with robot perception. The vast majority of existing work in this space limits robot perception to vision. We build perceptual models that use haptic, auditory, and proprioceptive data acquired through robot exploratory behaviors to go beyond vision. Our system learns to ground natural language words describing objects using

[2] View a video demonstrating the learning process on the Segbot at: `https://youtu.be/FL9IhJQOzb8`.

Table 2: Average Segbot survey responses from the two test groups and the proportion of task goals completed. Means in bold differ significantly ($p < 0.05$). Means in italics trend different ($p < 0.1$).

	Init Test	Trained Test
Survey Question	Likert [0-4]	
Tasks Easy	3.8	3.7
Robot Understood	1.6	**2.9**
Robot Frustrated	2.5	*1.5*
Use Navigation	2.8	2.5
Use Delivery	1.6	2.5
Goals Completed	Percent	
Navigation	90	90
Delivery	20	*60*

Figure 6: **Left**: the robot guesses an object described by a human participant as "silver, round, and empty." **Center**: a human participant guesses an object described by the robot as "light," "tall," and "tub." **Right**: objects used in the "I Spy" game divided into the four folds, from fold 0 on the left to fold 3 on the right.

supervision from an interactive human-robot "I Spy" game.

While corpora like ImageNet (Deng et al. [2009]) can provide a large set of labeled images to learn classifiers for words and noun phrases, properties like "heavy" are grounded in non-visual space. Annotating a similarly large body of objects with non-visual properties and gathering robot perception or even features (like weight) about them is costly and does not generalize across different robotic platforms. We propose the "I Spy" game as a paradigm to get a perceptual grounding system "off the ground" since it is fun for human users and requires less labor than straight annotation. This completed work provides a blueprint for perceptual grounding of the predicates "light" and "mug" from the earlier example (1). We later propose to continuously refine this bootstrapped perception in a fully integrated robotic system that uses dialog to clarify misunderstandings.

In this game, the human and robot take turns describing one object among several, then trying to guess which object the other has described (Figure 6 (Left, Center)). We demonstrate that our multi-modal system for grounding natural language outperforms a traditional, vision-only grounding framework by comparing the two on the "I Spy" task. Full details are available in Thomason et al. [2016].

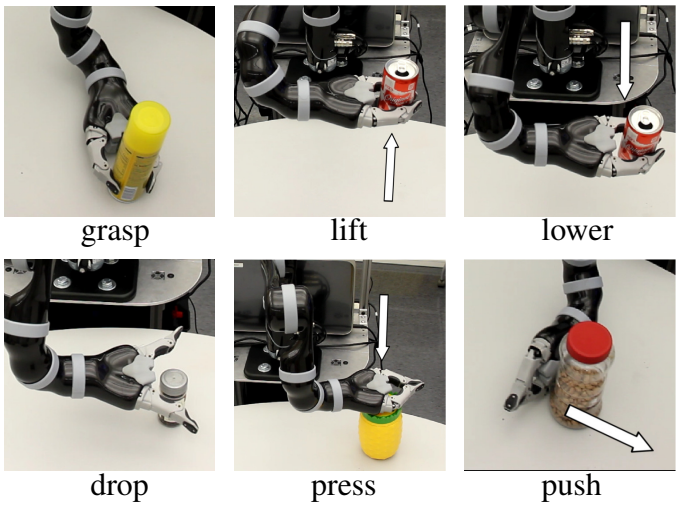


Figure 7: The behaviors the robot used to explore the objects. In addition, the *hold* behavior (not shown) was performed after the *lift* behavior by simply holding the object in place for half a second.

4.1 Methods

The robot used in this study was a Kinova MICO arm mounted on top of a custom-built mobile base which remained stationary during our experiment. The robot's perception included joint effort sensors in each of the robot arm's motors, a microphone mounted on the mobile base, and an Xtion ASUS Pro RGBD camera. The set of objects used consisted of 32 common household items including cups, bottles, cans, and other containers, shown in Figure 6 (Right). Some of the objects contained liquids or other contents (e.g., coffee beans) while others were empty. Contemporary work gives a more detailed description of this object dataset (Sinapov et al. [2016]), but we briefly describe the exploration and modalities below.

Prior to the experiment, the robot explored the objects using the methodology described by Sinapov et al. [2014a], and the dimensionality of the raw auditory, haptic, and proprioceptive data were reduced comparably (final dimensionality given in Table 3). In our case, the robot used 7 distinct actions: *grasp*, *lift*, *hold*, *lower*, *drop*, *push*, and *press*, shown in Figure 7. During the execution of each action, the robot recorded the sensory perceptions from *haptic* (i.e., joint efforts) and *auditory* sensory modalities. During the *grasp* action, the robot recorded *proprioceptive* (i.e., joint angular positions) sensory information from its fingers. The joint efforts and joint positions were recorded for all 6 joints at 15 Hz. The auditory sensory modality was represented as the Discrete Fourier Transform computed using 65 frequency bins.

In addition to the 7 interactive behaviors, the robot also performed the *look* action which produced three different kinds of sensory modalities: 1) an RGB color histogram of the object using 8 bins per channel; 2) Fast point feature histogram (*fpfh*) shape features (Rusu et al. [2009]) as implemented in the Point Cloud Library (Aldoma et al. [2012]); and 3) deep visual features from the 16-layer VGG network (Simonyan and Zisserman [2014]). The first two types of features were

Behavior	Modality		
	color	**fpfh**	**vgg**
look	64	308	4096
	audio	**haptics**	**proprioception**
grasp	100	60	20
drop, hold, lift, lower, press, push	100	60	

Table 3: The number of features extracted from each *context*, or combination of robot behavior and perceptual modality.

computed using the segmented point cloud of the object while the deep features were computed using the 2D image of the object.

Thus, each of the robot's 8 actions produced two to three different kinds of sensory signals. Each viable combination of an action and a sensory modality is a unique sensorimotor context. In our experiment, the set of contexts $\mathcal{C}$ was of size $2 \times 3 + 6 \times 2 = 18$. The robot performed its full sequence of exploratory actions on each object 5 different times (for the *look* behavior, the object was rotated to a new angle each time). Given a context $c \in \mathcal{C}$ and an object $i \in \mathcal{O}$, let the set $\mathcal{X}_i^c$ contain all five feature vectors observed with object i in context c.

For each language predicate p, a classifier G_p was learned to decide whether objects possessed the attribute denoted by p. This classifier was informed by context sub-classifiers that determined whether p held for subsets of an object's features.

The feature space of objects was partitioned by context. Each context classifier $M_c, c \in \mathcal{C}$ was a quadratic-kernel SVM trained with positive and negative labels for context feature vectors derived from the "I Spy" game. We defined $M_c(\mathcal{X}_i^c) \in [-1, 1]$ as the average classifier output over all observations for object $i \in \mathcal{O}$ (individual SVM decisions on observations were in $\{-1, 1\}$).

Following previous work in multi-modal exploration (Sinapov et al. [2014b]), for each context we calculated Cohen's Kappa $\kappa_c \in [0, 1]$ to measure the agreement across observations between the decisions of the M_c classifier and the ground truth labels from the "I Spy" game.[3] Given these context classifiers and associated κ confidences, we calculate an overall decision, $G_p(i)$, for $i \in \mathcal{O}$ for each behavior b and modality m as:

$$G_p(i) = \sum_{c \in \mathcal{C}} \kappa_c M_c(\mathcal{X}_i^c) \in [-1, 1] \tag{3}$$

The sign of $G_p(i)$ gives a decision on whether p applies to i with confidence $|G_p(i)|$.

For example, a classifier built for 'fat'$\in P$ could give $G_{\texttt{fat}}(\texttt{wide-yellow-cylinder}) = 0.137$, a positive classification, with $\kappa_{gr,au} = 0.515$ for the *gr*asp behavior's *au*ditory modality, the most confident context. This context could be useful for this predicate because the sound of the fingers' motors stop sooner for wider objects.

[3]We use κ instead of accuracy because it better handles skewed-class data than accuracy, which could be deceptively high for a classifier that always returns false for a low-frequency predicate. We round negative κ up to 0.

Language predicates and their positive/negative object labels were gathered through human-robot dialog during the "I Spy" game. The human participant and robot were seated at opposite ends of a small table. A set of 4 objects were placed on the table for both to see (Figure 6). We denote the set of objects on the table during a given game $\mathcal{O}_T$.

Human Turn. On the participant's turn, the robot asked him or her to pick an object and describe it in one phrase. We used a standard stopword list to strip out non-content words from the participant's description. The remaining words were treated as a set of language predicates, $\mathcal{H}_p$. The robot assigned scores S to each object $i \in \mathcal{O}_T$ on the table.

$$S(i) = \sum_{p \in \mathcal{H}_p} G_p(i) \tag{4}$$

The robot guessed objects in descending order by score (ties broken randomly) by pointing at them and asking whether it was correct. When the correct object was found, it was added as a positive training example for all predicates $p \in \mathcal{H}_p$ for use in future training.

Robot Turn. On the robot's turn, an object was chosen at random from those on the table. To describe the object, the robot scored the set of known predicates learned from previous play. Following Gricean principles (Grice [1975]), the robot attempted to describe the object with predicates that applied but did not ambiguously refer to other objects. We used a predicate score R that rewarded describing the chosen object i^* and penalized describing the other objects on the table.

$$R(p) = |O_T| G_p(i^*) - \sum_{j \in \mathcal{O}_T \setminus \{i^*\}} G_p(j) \tag{5}$$

The robot choose up to three highest scoring predicates $\hat{P}$ to describe object i^*, using fewer if $S < 0$ for those remaining. Once ready to guess, the participant touched objects until the robot confirmed that they had guessed the right one (i^*).

The robot then pointed to i^* and started a follow-up dialog in order to gather both positive and negative labels for i^*. In addition to predicates $\hat{P}$ used to describe the object, the robot selected up to $5 - |\hat{P}|$ additional predicates $\bar{P}$. $\bar{P}$ were selected randomly with $p \in P \setminus \hat{P}$ having a chance of inclusion proportional to $1 - |G_p(i^*)|$, such that classifiers with low confidence in whether or not p applied to i^* were more likely to be selected. The robot then asked the participant whether they would describe the object i^* using each $p \in \hat{P} \cup \bar{P}$. Responses to these questions provided additional positive/negative labels on object i^* for these predicates.

4.2 Experiments

In our "I Spy" task,[4] the human and robot take turns describing objects from among 4 on a tabletop using attributes (Figure 6). As an example, we suggested participants describe an object as "black rectangle" as opposed to "whiteboard eraser." Additionally, participants were told they could handle the objects physically before offering a description, but were not explicitly asked to use non-visual predicates. Once participants offered a description, the robot guessed candidate objects in order of computed confidence until one was confirmed correct by the participant.

[4]Video demonstrating the "I Spy" task and robot learning: `https://youtu.be/jLHzRXPCi_w`

In the second half of each round, the robot picked an object and then described it with up to three predicates. The participant was again able to pick up and physically handle objects before guessing. The robot confirmed or denied each participant guess until the correct object was chosen.

"I Spy" gameplay admits two metrics. The **robot guess** metric is the number of turns the robot took to guess what object the participant was describing. The **human guess** metric is the complement. Using these metrics, we compare the performance of two "I Spy" playing systems (**multi-modal** and **vision-only**). We also compare the agreement between both systems' predicate classifiers and human labels acquired during the game.

During the course of the game, the robot used its RGBD camera to detect the locations of the objects and subsequently detect whenever a human reached out and touched an object in response to the robot's turn. The robot could also reach out and point to an object when guessing.

To determine whether multi-modal perception helps a robot learn grounded language, we had two different systems play "I Spy" with 42 human participants. The baseline **vision only** system used only the *look* behavior when grounding language predicates. Our **multi-modal** system used the full suite of behaviors and associated haptic, proprioceptive, and auditory modalities shown in Table 3 when grounding language predicates.

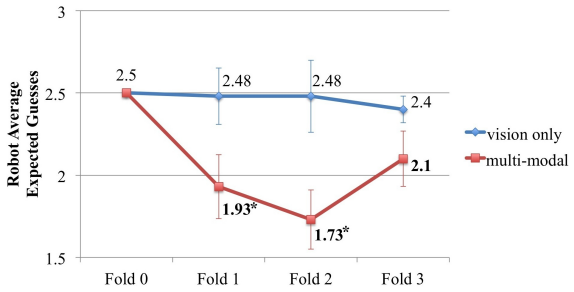
Data Folds. We divided our 32-object dataset into 4 folds. For each fold, at least 10 human participants played "I Spy" with both the **vision only** and **multi-modal** systems. Four games were played by each participant. The **vision only** system and **multi-modal** system were each used in 2 games, and these games' temporal order was randomized. Each system played with all 8 objects per fold, but the split into 2 groups of 4 and the order of objects on the table were randomized.

For fold 0, the systems were undifferentiated and so only one set of 2 games was played by each participant. For subsequent folds, the systems were incrementally trained using labels from previous folds only, such that the systems were always being tested against novel, unseen objects. This contrasts prior work using the "I Spy" game (Parde et al. [2015]), where the same objects were used during training and testing.

Human Participants. Our 42 participants were undergraduate and graduate students as well as some staff at our university. At the beginning of each trial, participants were shown an instructional video of one of the authors playing a single game of "I Spy" with the robot, then given a sheet of instructions about the game and how to communicate with the robot. In every game, participants took one turn and the robot took one turn. To avoid noise from automatic speech recognition, a study coordinator remained in the room and transcribed the participant's speech to the robot from a remote computer. This was done discretely and not revealed to the participant until debriefing when the games were over.

Results. To determine whether our **multi-modal** approach outperformed a traditional **vision only** approach, we measured the average number of robot guesses and human guesses in games played with each fold of objects. The systems were identical in fold 0 since both were untrained. In the end, we trained the systems on all available data to calculate predicate classifier agreement with human labels.

Robot guess. Figure 8 (Left) shows the average number of robot guesses for the games in each fold. Because we had access to the scores the robot assigned each object, we calculated the



Metric	System	
	vision only	**multi-modal**
precision	.250	.378+
recall	.179	.348*
F_1	.196	.354*

Figure 8: **Left:** Average expected number of guesses the robot made on each human turn with standard error bars shown. *Bold*: significantly lower than the average at fold 0 with $p < 0.05$ (unpaired Student's t-test). *: significantly lower than the competing system on this fold on participant-by-participant basis with $p < 0.05$ (paired Student's t-test). **Right:** Average performance of predicate classifiers used by the **vision only** and **multi-modal** systems in leave-one-object-out cross validation. *: significantly greater than competing system with $p < 0.05$. +: $p < 0.1$ (Student's un-paired t-test).

expected number of robot guesses for each turn. For example, if all 4 objects were tied for first, the expected number of robot guesses for that turn was 2.5, regardless of whether it got (un)lucky and picked the correct object (last)first.[5]

After training on just one fold, our **multi-modal** approach performs statistically significantly better than the expected number of turns for guessing (the strategy for the untrained fold 0 system) for the remainder of the games. The **vision only** system, by contrast, is never able to differentiate itself significantly from random guessing, even as more training data becomes available. We suspect the number of objects is too small for the **vision only** system to develop decent models of many predicates, whereas **multi-modal** exploration allows that system to extract more information per object.

Human guess. Neither the **vision only** nor **multi-modal** system's performance improves on this metric with statistical significance as more training data is seen. This result highlights the difficulty of the robot's turn in an "I Spy" framework, which requires not just good coverage of grounded words (as when figuring out what object the human is describing), but also high accuracy when using classifiers on new objects. Context classifiers with few examples could achieve confidence $\kappa = 1$, making the predicates they represented more likely to be chosen to describe objects. It is possible that the system would have performed better on this metric if the predicate scoring function R additionally favored predicates with many examples.

Predicate Agreement. Training the predicate classifiers using leave-one-out cross validation over objects, we calculated the average precision, recall, and F_1 scores of each against human predicate labels on the held-out object. Table 8 (Right) gives these metrics for the 74 predicates used by the systems.[6]

[5] 2.5 is the expected number for 4 tied objects because the probability of picking in any order is equal, so the expected turn to get the correct object is $\frac{1+2+3+4}{4} = \frac{10}{4} = 2.5$

[6] There were 53 predicates shared between the two systems. The results are similar for a paired t-test across these shared predicates with slightly reduced significance.

Across the objects our robot explored, our **multi-modal** system achieves consistently better agreement with human assignments of predicates to objects than does the **vision only** system.

Correlations to physical properties. To validate whether the systems learned non-visual properties of objects, for every predicate we calculated the Pearson's correlation r between its decision on each object and that object's measured weight, height, and width. As before, the decisions were made on held-out objects in leave-one-out cross validation. We found predicates for which $r > 0.5$ with $p < 0.05$ when the system had at least 10 objects with labels for the predicate on which to train.

The **vision only** system led to no predicates correlated against these physical object features.

The **multi-modal** system learned to ground predicates which correlate well to objects' height and weight. The "tall" predicate correlates with objects that are higher ($r = .521$), "small" ($r = -.665$) correlates with objects that are lighter, and "water" ($r = .814$) correlates with objects that are heavier. The latter is likely from objects described as "water bottle", which, in our dataset, are mostly filled either half-way or totally and thus heavier. There is also a spurious correlation between "blue" and weight ($r = .549$). This highlights the value of multi-modal grounding, since words like "half-full" cannot be evaluated with vision alone when dealing with closed containers that have unobservable contents.

5 Multi-Modal Word Synset Induction

Words in natural language can be polysemous (a single word with multiple meanings) as well as synonymous (multiple words with the same meaning). Word sense induction attempts to determine the senses of a polysemous word type by clustering the contexts in which it occurs.

Most prior work in this task has used *linguistic* context to determine senses. Some works use *visual* context to ground senses in perceptual data, and still others use both. We go beyond word sense induction to the task of word sense synonym set (*synset*) induction. Given a noun phrase and an associated set of images with textual context, we perform polysemy and synonymy detection through clustering to obtain synsets. We evaluate our approach by measuring how well we recover membership in gold standard sets of senses in ImageNet (Deng et al. [2009]). We find that polysemy detection improves precision metrics, while synonymy detection improves recall metrics. When polysemy and synonymy detection are chained, a mixture of both textual and visual features of word observations gives better hypothesized synsets than textual or visual features alone.

We first perform WSI, detecting polysemy in words to form senses, and then group induced senses together into synsets. Such synsets could be used in WSD tasks, and additional information about each word sense is made available by grouping it with other, synonymous senses.

Understanding a word or phrase can be done in both textual and visual space. For instance, the two readings of "bat" are both textually and visually distinct. When modeling polysemy and synonymy, we perform *multi-modal synset induction*, considering both textual and visual contexts for words.

Resources like ImageNet contain visual representations of synsets. However, ImageNet required extensive manual annotation from humans to construct, is limited to its current coverage, and is and only available in English. Our methodology enables the automatic construction of an

ImageNet-like resource from a collection of images and associated texts, and could be employed for domain-specific or non-English synset induction. This completed work establishes a methodology for resolving the ambiguity of "light" in (1).

5.1 Methods

We select a subset of synsets from ImageNet that: 1) are leaves in the WordNet hierarchy (*i.e.* words with no hyponyms); 2) are not used to train the VGG network (Simonyan and Zisserman [2014]); 3) have associated images; 4) are embedded in pages discovered through reverse image search with text; and 5) together represent polysemous and synonymous noun phrases as well as those that were neither. We note that (5) makes our task strictly harder than structured word sense disambiguation tasks such as Manandhar et al. [2010] and some previous word sense induction datasets (Loeff et al. [2006], Saenko and Darrell [2008]), because not all our noun phrases have multiple senses. Table 4 gives the number of noun phrases which participated in polysemous and synonymous relationships among the 6,710 synsets S we extracted from ImageNet.

syn	**poly**	**both**	**neither**
4019	804	1017	2586

Table 4: Number of noun phrases that are synonymous, polysemous, both, or neither in the subset of ImageNet synsets we consider.

For each image in each synset, we extract visual features, perform reverse image search to find and extract web text from pages in which the image is embedded, and extract textual features from those pages. We deconstruct these gold-standard synsets to associate images directly with noun phrases, making a monosemous (single-sense per word) baseline. We use polysemy-detecting WSI followed by synonymy detection to reconstruct synsets from raw images and associated text, first breaking noun phrases into multiple senses, then clustering those senses to form synsets.

Specifically, we associate individual noun phrases with subsets of the images/texts in the ImageNet synsets to which they belong, mixing together multiple senses. In this way, the noun phrase "kiwi" is associated with images and texts of both the fruit and the bird. Figure 9 demonstrates this conversion from ImageNet synsets to mixed-sense noun phrases. Our task is then to recover the synsets by detecting clusters of sense meaning in textual and visual space, such as the three sets of images representing meanings of "kiwi" (polysemy detection, illustrated in Figure 10 (Left)), then merging those senses together to form hypothesized synsets (synonymy detection, illustrated in Figure 10 (Right)). This induces not just word senses but whole synsets from observed images and text associated with noun phrases.

We take the synsets V used to train the VGG network as training data. For a subset of the training images, we performed reverse image search and scraped the text of webpages on which those images appeared. We performed latent semantic analysis (LSA) on *tf-idf* vectors of bag-of-words representations of this text to create a 256-dimensional latent textual feature space.

For each synset $s \in S$, we downloaded up to 100 images per noun phrase associated with s in ImageNet and extracted deep visual features as the activations of the 4,096-dimensional, penulti-

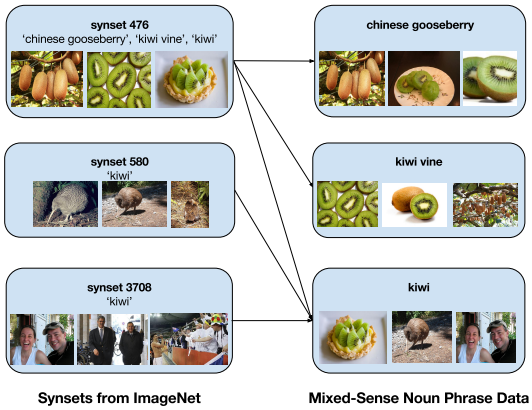


Figure 9: ImageNet synsets are composed of noun phrases, each of which may describe multiple synsets. We divide the images of a synset randomly among noun phrases, then collapse all images associated with each noun phrase to form an inventory of mixed sense noun phrases.

mate later of the 16-layer VGG network (Simonyan and Zisserman [2014]), giving a set of image observations I_s. For each image, we performed reverse image search to gather a corpus of text from web pages on which that image was embedded, then created a textual feature representation for the image by embedding the text of those pages as a single document in our LSA space. This gives a set of textual observations T_s parallel to I_s for each synset s.

For each synset s, we have observations $O_s = \langle I_s, T_s \rangle$. Every synset s is composed of noun phrases. We denote the union over all such noun phrases for all $s \in S$ as NP. For each noun phrase $np \in NP$ in the dataset, we associated image observations with np from each synset in which it participated by dividing each O_s evenly among participating noun phrase (Figure 9). We refer to noun phrase observations as O_{np}.

After this process, O_{np} for a polysemous np contains observations from multiple senses. Additionally, the observations O_s for a synset with multiple synonymous noun phrases have been split across those noun phrases. Our task is to try to recover the original O_s observations at the synset level starting with the mixed-sense O_{np} observations at the noun phrase level. To do this, we perform multi-modal synset induction.

Discovering polysemy in noun phrase np in this context is equivalent to dividing O_{np} into k disjoint sets $O_{np,k}$ representing k senses of np (Figure 10 (Left)). Given these senses, discovering synonymy among noun phrase senses involves joining O_{np_i,k_i} and O_{np_j,k_j} noun phrase senses that refer to the same concept (e.g. "Chinese gooseberry" and "kiwi" the fruit) to form final, reconstructed synsets $r \in R$ for R the set of hypothesized synsets (Figure 10 (Right)).

Our goal is to reconstruct the original synsets for our subset of ImageNet using the perceptual and textual information extracted from images now associated with potentially ambiguous noun phrases. We performed reconstructions using only textual features, only visual features, and an equally-weighted representation of both textual and visual features. Our reconstruction effort takes place in two phases: polysemy detection and synonymy detection. We evaluation our reconstruction using two metrics: the *v-measure* and the *paired f-measure*, both used in the SemEval-2010

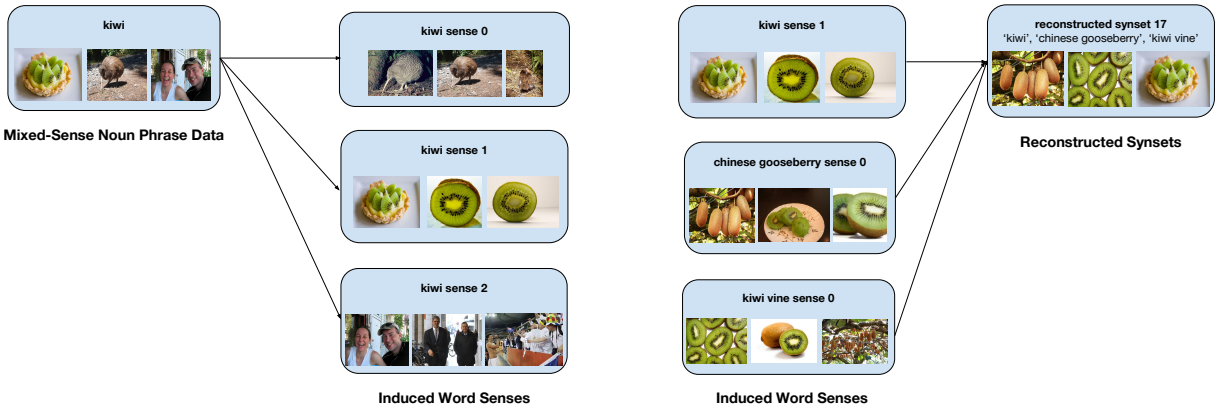


Figure 10: **Left:** In polysemy detection, clustering within each mixed noun phrase sense yields some number of induced senses per noun phrase. **Right:** In synonymy detection, clustering is performed over the induced word senses produced during polysemy detection. Senses are clustered together to construct synsets.

Word Sense Induction and Disambiguation task (Manandhar et al. [2010]).

Polysemy detection. We use a non-parametric k-means algorithm that selects a number of clusters k for each set of noun phrase observations O_{np} based on the gap statistic (Tibshirani et al. [2001]).

The *gap* of a clustering of data D with k clusters is the average difference between the within-dispersion of clusterings over uniform data and the within-dispersion of the clustering of D with k clusters. Reference datasets B are generated with points uniformly distributed within the range of the features in D. k-means clustering is done for each $b \in B$ and the within-dispersion measure $W_{k,b}$ is calculated as

$$W_{k,b} = \sum_{r=1}^{k} \frac{D_r}{2n_r}$$

with D_r the sum of pairwise distances for all points in cluster r and n_r the number of points belonging to cluster r in reference dataset b. Data D is then clustered into k clusters and W_k is calculated similarly on the result. We use cosine distance to measure the space between points and generate $|B| = 100$ reference datasets for each set of input data D. The gap at k clusters is then calculated as

$$gap(k) = \frac{1}{B} \sum_{b \in B} log(W_{kb} - log(W_k))$$

Given this statistic, the optimal $k*$ is estimated as the smallest k such that

$$gap(k) \geq gap(k+1) - s_{k+1}$$

For $s_{k+1} = s_k \sqrt{1 + \frac{1}{|B|}}$ and s_k the standard deviation of the $W_{k,b}$ values for $b \in B$. Intuitively, by selecting $k*$ in this way, we get the largest number of clusters that reduce the within-dispersion of the clustering by more than chance (see Tibshirani et al. [2001] for more details).

Additionally, we enforce the constraint that no induced sense (cluster of observations) has fewer than 20 observations. This constraint is estimated from the training set V, which contains an average of 131.24 observations per cluster with standard deviation 110.86, so that less than 20 observations per cluster is more than one standard deviation away from the expected mean. This constraint is applied post-hoc, merging degenerate clusters into their nearest neighbor.

Thus, for a given noun phrase n, we cluster the observations O_{np} into $k*$ senses, using the gap statistic above to determine the number of clusters $k*$. This gives us observation sets O_{np,k_i} for $k_i \in [0, k*)$. Together, all these observation sets form a set of induced senses G.

Synonymy detection. In synonymy detection, for each induced sense in G we compute a mean m to form a collection of mean vectors M. We then compute the pairwise cosine distance between all mean vectors and perform greedy merges of the nearest means to form synsets R.

We continue merging clusters until an estimated K synsets is reached. We estimate K from the average number of word senses per synset in the training set V. We enforce a constraint that no synset contain more than 32 distinct word senses from G. This constraint is also estimated from training data V, where 32 was the maximum number of word senses in a single synset.

Membership in each final synset $r \in R$ is the union of observations of the senses $g \in G$ whose observations were merged (e.g. $r = \cup_i g_i$).

When discovering polysemy, we want the minimum number of word senses that explain the instances we observe, while in synonymy we only want to join together highly similar senses into synsets. We note that using the gap statistic to estimate an optimal number of clusters for synonymy detection would be inappropriate because we know k^* is on the order of $|G|$. That is, we know the number of synsets is closer to the number of word senses than to 1. The gap statistic is best applied when looking for a minimum k^*, but further sensible divisions of k^* well-separated clusters may exist contained within larger clusters (Tibshirani et al. [2001]).

Evaluation via v-measure. The *v-measure* (Rosenberg and Hirschberg [2007]) of a reconstructed set of synsets is the harmonic mean of its *homogeneity* and *completeness* with respect to the gold synsets. High homogeneity means the reconstructed synsets mostly contain observations that correspond to a single gold synset, while high completeness means each gold synset's observations are assigned to the same reconstructed synset. These are defined in terms of the class entropies $H(S)$ and $H(R)$ of the gold-standard ImageNet synsets S and hypothesis reconstructed synsets R and their conditional entropies $H(S|R)$ and $H(R|S)$. Specifically, homogeneity $h(S, R)$ is calculated

as follows:

$$H(S) = -\sum_{i=1}^{|S|} \frac{\sum_{j=1}^{|R|} a_{ij}}{N} \log \frac{\sum_{j=1}^{|R|} a_{ij}}{N}$$

$$H(S|R) = -\sum_{j=1}^{|R|} \sum_{i=1}^{|S|} \frac{a_{ij}}{N} \log \frac{a_{ij}}{\sum_{k=1}^{|S|} a_{kj}}$$

$$h(S,R) = \begin{cases} 1 & H(S) = 0 \\ 1 - \frac{H(S|R)}{H(S)} & H(S) > 0 \end{cases}$$

with a_{ij} the number of observations of gold synset S_i that ended up in hypothesized synset R_j and N the total number of observations in the dataset. Completeness $c(S,R)$ is defined as follows:

$$H(R) = -\sum_{j=1}^{|R|} \frac{\sum_{i=1}^{|S|} a_{ij}}{N} \log \frac{\sum_{i=1}^{|S|} a_{ij}}{N}$$

$$H(R|S) = -\sum_{i=1}^{|S|} \sum_{j=1}^{|R|} \frac{a_{ij}}{N} \log \frac{a_{ij}}{\sum_{k=1}^{|R|} a_{ik}}$$

$$c(S,R) = \begin{cases} 1 & H(R) = 0 \\ 1 - \frac{H(R|S)}{H(R)} & H(R) > 0 \end{cases}$$

The v-measure is then calculated as follows:

$$v(S,R) = \frac{2 * h(S,R) * c(S,R)}{h(S,R) + c(S,R)}$$

Evaluation via paired f-measure. The *paired f-measure* is the harmonic mean of the paired precision and recall between the gold synsets and the hypothesized reconstructions. Rather than count membership overlap between two sets, paired f-measure allows us to compare membership overlap between sets of sets.

Specifically, we count the number of observation pairs (o_i, o_j) that are members of both synset s and hypothesized synset r to get an overlap score between each $s \in S$ and $r \in R$. There are $\binom{|s|}{2}$ observation pairs for each s and $\binom{|r|}{2}$ observation pairs for each r, across all such s and r comprising $C(S)$ gold pairs and $C(R)$ reconstructed pairs, respectively. Then paired precision $p(S,R)$, recall $r(S,R)$, and f-measure $f(S,R)$ is defined as

$$p(S,R) = \frac{|C(S) \cap C(R)|}{|C(R)|}$$

$$r(S,R) = \frac{|C(S) \cap C(R)|}{|C(S)|}$$

$$f(S,R) = \frac{2 * p(S,R) * r(S,R)}{p(S,R) + r(S,R)}$$

5.2 Experiments

We consider several conditions when reconstructing synsets for our task. We perform polysemy and synonymy detection with textual features only, visual features only, and an equal weight of textual and visual features given when calculating distance between observations. We examine the effects of performing only polysemy detection (equivalent to many past WSI works) as well as polysemy followed by synonymy detection.

Quantitative Results. Table 5 shows the results for the polysemy-detecting WSI step. We note that across all modalities, the homogeneity and paired precision increase after performing polysemy detection to split noun phrases into distinct senses.

features	k	h	c	v	p	r	f
monosemous	7922	0.968	0.943	0.955	0.694	0.519	0.594
text	11535	$*$0.980	0.905	0.941	$*$0.725	0.381	0.499
vis	18913	$*$0.991	0.853	0.917	$*$0.839	0.209	0.335
text+vis	15395	$*$0.975	0.874	0.922	$*$0.803	0.275	0.409

Table 5: Number of hypothesized word senses (**k**), homogeneity (**h**), completeness (**c**), v-measure (**v**), paired precision (**p**), recall (**r**), and f-measure (**f**) of our induced word senses with textual features only, visual features only, and both when performing polysemy detection. The **monosemous** synsets are the sense-unaware baseline of mixed-sense noun phrases treated as monosemous (e.g. one sense of "kiwi" which contains instances of people, fruits, and birds). $*$: higher than corresponding baseline.

Table 6 shows the results for the synonymy step constructs hypothesis synsets R. We give results both for detecting synonymy in each modality given the polysemy detection results in that modality as well as given the "gold", perfectly-detected word senses. Giving the gold senses as input allows us to examine the synonymy step's performance in isolation, while giving the output of polysemy detection gives us the performance of the entire pipeline.

We note that given the gold senses, synonymy detection across all modalities improves the completeness and paired recall of the hypothesized synsets. Additionally, the combined text plus vision full pipeline achieves the highest v-measure and paired f-measure compared to the uni-modal pipelines, though this does not hold if the synonymy step starts with imperfect induced senses. This shows that the multi-modal pipeline steps work together, with the synonymy step overcoming some errors produced in the polysemy step, more effectively than the uni-modal pipelines.

Qualitative Results. We look at the reconstructed synsets for the multi-modal text plus vision pipeline. Returning to our "kiwi" example, we find the hypothesized synsets containing instances from "kiwi", shown in Figures 11. The reconstructed sets appear to correspond to: the whole fruit; the bird; close-up pictures of the cut-open fruit; the fruit cut in half for eating; and people.

$\|G\|$	features	k	h	c	v	p	r	f
gold $\|G\|$		9755	1.0	0.945	0.972	1.0	0.519	0.683
gold	text	5476	0.895	$*$0.966	0.929	0.259	$*$0.722	0.382
gold	vis	5476	0.941	$*$0.997	0.968	0.473	$*$0.909	0.622
gold	text+vis	5476	0.936	$*$0.995	0.965	0.441	$*$0.904	0.593
monosemous		7922	0.968	0.943	0.955	0.694	0.519	0.594
text	text	6476	0.855	0.909	0.881	0.173	0.482	0.255
vis	vis	10618	0.904	0.875	0.889	0.334	0.343	0.339
text+vis	text+vis	8643	0.893	0.900	0.896	0.290	0.456	0.354

Table 6: The metrics and synset evaluation sets of Table 5 for synonymy detection over resulting synsets, with **k** now the number of hypothesized synsets. The senses in $|G|$ were clustered using the **features** listed, either textual only, visual only, or both. The **gold** set $|G|$ is a perfect word-sense induction from the baseline noun phrases. It represents the upper bound of the polysemy detection step and thus the optimal input for the synonymy detection step. $*$: higher than corresponding baseline.

The "people" set is particularly interesting and highlights why evaluating with ImageNet as a gold standard makes performance gains difficult. It is unreasonable to expect any algorithm to be able to distinguish, for example, photos of Croatian from Ukranian peoples, or people who could be described as "energizers" from those that could not. The inclusion of pictorial representations of humans described as "energizers' or "inferior' in ImageNet is noise any algorithm for synset induction will have difficulty overcoming. The reconstruction method instead grouped all these senses of noun phrases referring to people together in one large synset. We suspect that a human evaluation of ImageNet synsets versus our reconstructed ones would find that ours were at least as sensible as ImageNet's, if not more, and plan this evaluation as proposed work.

Comparing to the uni-modal pipelines, using text alone produced two senses for "kiwi": the fruit (correctly synonymous with "kiwi vine" and "Chinese gooseberry"); and the bird and person mixed together with a sense of "pen" (baby pen). Using vision alone produced two monosemous synsets for the bird sense; membership in a humans synset (including "creole" and "preceptor"); and five distinct fruit senses, two correctly synonymous with "kiwi vine"/"Chinese gooseberry", and another mistakenly combined with "honeydew melon" and "Persian melon" due to visual similarity.

A multi-modal approach in which visual and textual features are considered together outperforms uni-modal approaches on the combined task of polysemy detection followed by synonymy detection on the induced (imperfect) senses.

6 Short-Term Proposed Work

Our existing works on improving semantic parsing over time through dialog (Section 3), grounding natural language predicates in robot perception (Section 4), and inducing synsets from words

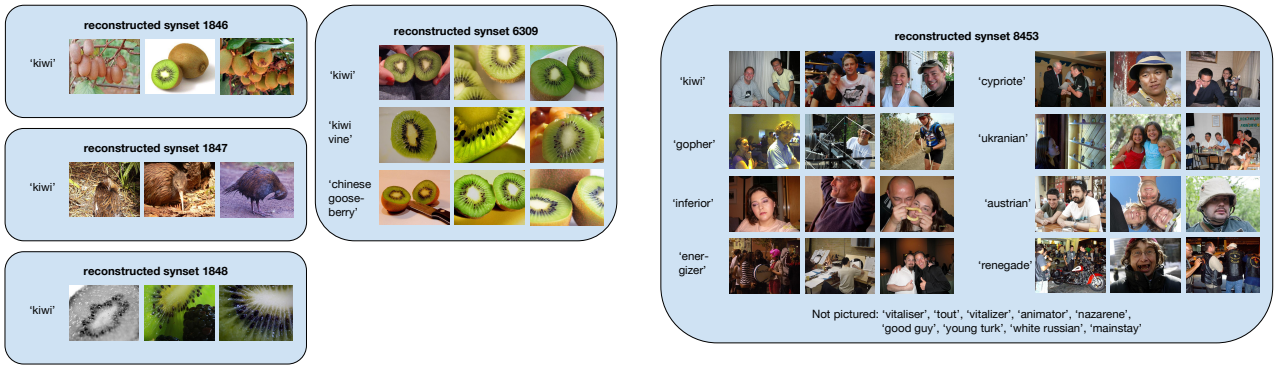


Figure 11: Our multi-modal pipeline's hypothesized synsets containing the observations labeled with the noun phrase "kiwi". Observation images are shown alongside the noun phrase label they had in the baseline, monosemous system on which we performed polysemy and synonymy detection to arrive at these synsets. We note that "gopher' appears in WordNet as a native Minnesotan. The words "energizer' (synonymous with "vitaliz(s)er" and "animator") and "inferior" appear as adjectives in WordNet that can describe people.

associated with observations (Section 5) run in orthogonal, but complimentary, directions. All completed work points towards an integrated, embodied system comprised of dialog, semantic parsing, and perception. Our short term proposed work primarily aims to realize this integrated system, enabling a robot to understand (1) while simultaneously improving its parsing and perception components through the interaction with the human interlocutor.

We also intend to produce synsets from the natural language predicates used to describe objects in Section 4, using methodology inspired by Section 5. Forming synsets from these multi-modal predicates should allow us to tease out polysemous predicates senses such as those for "light", as well as build stronger classifiers by combining the data of related predicates, such as "round" and "cylindrical".

Given predicate synset classifiers, we can develop a fully integrated robotic system that consists of a dialog agent, a semantic parser, a knowledge base, and a multi-modal perceptual grounding system. This integration should allow both parser strengthening through dialog supervision, as in Section 3, semantic re-ranking via perceptual confidence, and perception strengthening through dialog-based feedback.

6.1 Synset Induction for Multi-modal Grounded Predicates

In our experiments with multi-modal, grounding linguistic semantics (Section 4), we discovered that people use some polysemous words (e.g. "light") as well as effectively synonymous words (e.g. "round" and "cylindrical") when describing objects. By applying a synset induction algorithm to the predicates learned from the "I, Spy" game of Thomason et al. [2016], we could tease apart polysemous word senses and strengthen perceptual classifiers by combining synonymous predicates' data.

This kind of learning would be helpful in a deployed system partly because it can learn domain-specific polysemy and synonymy based on data. For example, in the office domain, the command "Fetch me a pen" may mean a robot should bring a writing pen or that it should bring a whiteboard pen. The polysemy step should be able to separate this domain senses of "pen". The synonymy step should subsequently merge the paper-relevant sense with a sense of "pencil" and the whiteboard-relevant sense with a sense of "marker".

In completed work, every observation is associated with a unique label (Thomason and Mooney [2016]). This means each pair of an image and the text of webpages it appeared on was associated with a single noun phrase, such as "kiwi" or "Chinese grapefruit". However, in the "I, Spy" game, and in robot perception in general, an observation can be associated wtih many labels. For example, the same object may be described as "blue", "cylindrical", and "bottle". A baseline for this proposed work would simply duplicate each object to exist as an instance for every label and run the synset induction algorithm as it exists. However, because perceptual contexts offer many more than the two modalities observed in completed work, there are certainly more interesting ways to frame the problem with multi-modality and multi-label objects.

6.2 Grounding Semantic Parses against Knowledge and Perception

In completed work, we trained a semantic parsing system with only bootstrapping initial data by using dialog with human users as passive supervision to generate more parser training data (Section 3). The parses produced were grounded against a knowledge base of facts regarding the surrounding office space. Knowledge base predicates in the semantic parser required querying against this information, while other types of predicates (such as `and`) could be handled logically. Proposed work would augment these parses with *perceptual* predicates grounded not by querying the knowledge base but by consulting sensory information and learned perceptual synset classifiers.

In this way, a person could specify a command like "Bring me a black eraser from Peter's office". The knowledge base predicate `owns` is still necessary to resolve "Peter's office", while perceptual predicates `black` and `eraser` will require the robot to evaluate items in the target office that satisfy these criteria. Several challenges need to be overcome to accomplish this integrated system. However, the integration also allows for novel information feedback loops.

Predicate Induction. In basic semantic parsing, the ontology of predicates natural language words can map into is fixed. When a new word is seen during parser training, part of the parser training task is to find the semantic form composed of these fixed ontological predicates that matches the word. In a system with perceptual predicates, however, new words may not mean something in the existing ontology, some new perceptual classifier that needs to be learned.

Thus, we propose to perform *predicate induction* during parsing on some new words that behave like perceptual predicates. Synset induction, as described in Section 6.1, could be performed intermittently on predicate classifiers to collapse some new induced predicates into existing classifiers (synonymy) as well as flag some words as polysemous during parsing since they map to two distinct perceptual contexts. The parsing step will be relieved of synonymy and polysemy detection for words describing perceptual predicates (because this will be handled by synset induction),

but must still identify words that trigger predicate induction (for example, finding those that are behaving as adjectives and nouns).

Semantic Re-ranking from Perception Confidence. A user utterance can be parsed into many candidate semantic forms. Probabilistic semantic parsers, such as the one used in our completed work, attempt to produce the best form first based on learned heuristics. The first form produced is not always correct, but the correct form may be found in some k-length beam from the parser.

We propose to maintain a list of k candidate forms for an utterance from the semantic parser. When evaluating explorable objects, the robot can use perceptual classifiers from all candidate forms and decide whether some object satisfies the constraints of a given form. Confidence in whether a candidate object is correct will be a combination of the parser's confidence in a form and perception's confidence in that form's predicates' application to the candidate. Figure 12 demonstrates this for "the light mug" and two candidate objects, where the predicates the parser is most confident in are not the ones corroborated by the environment, but re-ranking corrects this.

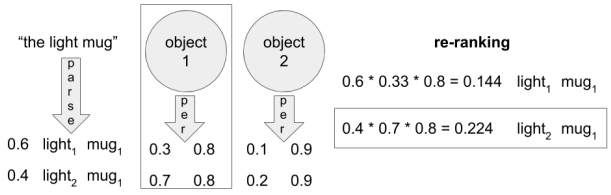


Figure 12: Candidate parses of "the light mug" contain the two senses of "light" (color and weight). Only the second sense is corroborated by the available objects in the environment, and the highest combined score (parser + perception classifiers) achievable is for the second, less confident parse together with object 1.

After confirming with the requesting human that the correct object was delivered, the semantic form that triggered the choice can be paired with the utterance to form a new training example for the parser, in the fashion of completed work (Section 3).

Perception Training Data from Dialog. In addition to the opportunities for parser training data induction from dialog after finding a target object, an integrated system can provide feedback from dialog to perception. Once a target object for some action has been identified and confirmed by a human user as correct, that object can be added as a positive example for the predicates used to describe it in the semantic form chosen. Thus, in the above example, once the right mug is identified, the perceptual classifiers for "mug", "green", and the color-sense of "light" gain an additional positive example—the mug brought to the person.

6.3 Related Works in Progress.

Two directions of proposed work are currently being explored primarily by colleagues of the author. These methods could be incorporated into the fully integrated robotic system.

Improving Speech Recognition through Parsing. Completed work uses a text-based interface between the user and embodied agent. A more natural interaction would allow human users to verbalize commands and responses to the agent. However, off-the-shelf speech recognition systems typically have high word-error rates unless trained extensively on individual user voices. Even small word-error rates can render utterances unparseable.

Work is under progress to add a speech recognition layer to our integrated system that utilizes the parser to improve speech recognition accuracy (Rodriquez et al. [2016]). Given an utterance, a speech recognizer can produce a beam of k candidate transcriptions. Each transcription can be run through the semantic parser. The transcription that is both parseable and gives the has the highest weighted confidence score between the speech recognizer and parser will be chosen as the correct transcription. This transcription can be paired with the audio signal to generate a new training pair for the speech recognizer. In this way, the speech recognition module can be improved over time as the embodied agent converses with human users.

Learning Dialog Policy. Completed work uses a static, hand-coded dialog policy to resolve confusion the agent has when taking commands from human users. This policy centers around slot-filling for commands and their known arguments, done by estimating a discrete belief about the user goal from an otherwise continuous one.

Work to replace this with a POMDP-based policy (Young et al. [2013]) that considers a continuous belief state about the user goal is currently under review (Padmakumar et al. [2016]). Such a policy should converge to user goals more quickly by accurately taking an expected shortest dialog path from the current understanding of the user's goal to a complete and confident understanding.

7 Long-Term Proposed Work

While our short-term proposals focus on the realization of an integrated system, these longer-term proposals aim to improve various aspects of that system once it exists. Each of these proposals is orthogonal and complementary, such that the implementation of each should improve the performance of the system as a whole.

Intelligent Exploration of Novel Objects. In the proposed integrated system, every exploration behavior is performed on novel objects in order to build a complete perceptual feature representation. These behaviors are slow to perform and cause wear on the robotic arm. If a person has requested "a pink marker", performing the full suite of exploratory behaviors is also unnecessary.

We propose to explore novel objects using only the behaviors necessary to ground perceptual predicates appearing in a given user utterance. Thus, for "a pink marker", the system should consult the predicate classifiers for "pink" and "marker" and identify the perceptual contexts (discussed in Section 4) that provide the most information to these classifiers. In this example, the *look* behavior followed by relevant visual feature extraction is likely sufficient to identify whether an object can be described with both "pink" and "marker". By contrast, asking for "the heavy mug" would require the robot to *lift* candidate objects.

Positive-unlabeled Learning for Perception. In completed work on perception (Section 4), perceptual classifiers are implemented as linear combinations of decisions from support vector machines operating in each perceptual context. SVMs draw a decision boundary in a feature space given labeled positive and negative examples. We add a stage to the "I, Spy" game where we explicitly ask human users whether predicates apply to a particular object. This allows us to gather negative examples alongside the natural positive examples that arise from identifying a described object.

We propose to replace the SVMs with positive-unlabeled classifier methods (Liu et al. [2003], Elkan and Noto [2008]). Some of these methods include streaming information, more like the additional examples our system will receive over time from conversational feedback (Chang et al. [2016]). In this way, when building classifiers for concepts like "heavy", we will only need positive examples of heavy objects. This will remove the need for gathering negative examples during dialog with humans.

Leveraging Accommodation. In most natural language understanding work, understanding user utterances is done by tweaking algorithms in the system to better understand what is spoken. However, the overarching goal of an NLU is often to communicate effectively with a human partner. Effective communication can come from more robust understanding, but it can also come from better input utterances from the human partner. In particular, if a user adapts her speech so that the system can better understand, effective communication is still achieved. In a vanilla dialog system, the user has no way to know what words or syntactic constructions the system understands best.

Past work by the author has explored the connection between user learning and tutoring dialog success (Thomason et al. [2013]) as it relates to accommodation. Accommodation is a conversational phenomenon in which interlocutors converge to shared referring expressions, lexical and syntactic choices, cadence, volume, and other vocal variations (Lakin et al. [2003], Gravano et al. [2015], Lubold et al. [2015]).

We propose to leverage accommodation to improve natural language understanding in our integrated system. When responding to a human user, rather than using template-based conversational responses, our dialog agent will consider a range of possible utterances and rank those utterances based on how well the semantic parser can understand them. Previous work has used a similar strategy to influence lexical choices (Lopes et al. [2013]). Through accommodation, we can expect human users to adopt the lexical and syntactic choices of the robotic agent as the conversation proceeds. By choosing to speak an utterance that the semantic parser understands most easily to the user, the system tacitly encourages the user to make lexical choices the parser itself will better understand.

8 Conclusion

For humans and robots to communicate effectively in shared environments like homes, offices, and factories, robots must be able to understand and respond to natural language from humans. We present completed work on using semantic parsing together with dialog to bootstrap and iteratively improve language understanding through conversations with humans, learning to ground predicates

like "light" in multi-modal perceptual space for a robot with an arm, and inducing word sense synonym sets.

We propose to improve predicate grounding by word sense synonym set induction to address ambiguous predicates and improve recognition performance for rare predicates with more common synonyms. We propose to then integrate semantic parsing, dialog, and perception into an embodied robotic system that improves its semantic and perceptual understandings over time with passive supervision from dialog with human users.

We then cover a series of orthogonal directions for future work to improve the overall integrated robotic system. These include intelligently exploring new objects and using positive-unlabeled learning for perceptual grounding, and leveraging accommodation to improve semantic recognition.

References

E. Bruni A. Lazaridou and M. Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL 2014 (52nd Annual Meeting of the Association for Computational Linguistics)*, pages 1403–1414, 2014.

Aitor Aldoma, Zoltan-Csaba Marton, Federico Tombari, Walter Wohlkinger, Christian Potthast, Bernhard Zeisl, Radu Bogdan Rusu, Suat Gedikli, and Markus Vincze. Point cloud library. *IEEE Robotics & Automation Magazine*, 1070(9932/12), 2012.

Yoav Artzi and Luke Zettlemoyer. Bootstrapping semantic parsers from conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 421–432, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Yoav Artzi and Luke Zettlemoyer. UW SPF: The University of Washington Semantic Parsing Framework, 2013a.

Yoav Artzi and Luke Zettlemoyer. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1(1): 49–62, 2013b.

Kobus Barnard and Matthew Johnson. Word sense disambiguation with pictures. *Artificial Intelligence*, 167:13–30, 2005.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. Textual inference and meaning representation in human robot interaction. In *Joint Symposium on Semantic Processing*, Trento, Italy, 2013.

J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic parsing on Freebase from question-answer pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Stefan Bordag. Word sense induction: Triplet-based clusering and automatic evaluation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 137–144, 2006.

Shiyu Chang, Yang Zhang, Jiliang Tang, Dawei Yin, Yi Chang, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Positive-unlabled learning in streaming networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016)*, 2016.

Xinlei Chen, Alan Ritter, Abhinav Gupta, and Tom Mitchell. Sense Discovery via Co-Clustering on Images and Text. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition*, 2009.

Antonio Di Marco and Roberto Navigli. Clustering and diversifying web search results with graph-based word sense induction. *Computational Linguistics*, 39(3):709–754, 2013.

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

Kais Dukes. Semeval-2014 task 6: Supervised semantic parsing of robotic spatial commands. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 45–53, 2014.

C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pages 213–220, 2008.

Christiane D. Fellbaum. *WordNet: An Electronic Lexical Database*. MITP, Cambridge, MA, 1998.

Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

Agustın Gravano, Štefan Benuš, Rivka Levitan, and Julia Hirschberg. Backward mimicry and forward influence in prosodic contour choice in standard american english. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

H. Paul Grice. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York, 1975.

Sergio Guadarrama, Erik Rodner, Kate Saenko, and Trevor Darrell. Understanding object descriptions in robotics by open-vocabulary object retrieval and detection. *International Journal of Robotics Research (IJRR)*, 35(1-3):265–280, 2015.

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

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Piyush Khandelwal, Fangkai Yang, Matteo Leonetti, Vladimir Lifschitz, and Peter Stone. Planning in Action Language $\mathcal{BC}$ while Learning Action Costs for Mobile Robots. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, 2014.

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

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE International Conference on Human-robot Interaction*, HRI '10, pages 259–266, 2010.

Thomas Kollar, Jayant Krishnamurthy, and Grant Strimel. Toward interactive grounded language acquisition. In *Robotics: Science and Systems*, 2013.

Evan Krause, Michael Zillich, Thomas Williams, and Matthias Scheutz. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.

H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, 2014.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1545–1556, 2013.

Jessica L. Lakin, Valerie E. Jefferis, Clara Michelle Cheng, and Tanya L. Chartrand. The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of Nonverbal Behavior*, 27(3):145–162, 2003. ISSN 1573-3653. doi: 10.1023/A:1025389814290.

Joohyung Lee, Vladimir Lifschitz, and Fangkai Yang. Action Language $\mathcal{BC}$: A Preliminary Report. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

P. Liang, M. I. Jordan, and D. Klein. Learning semantic correspondences with less supervision. In *Association for Computational Linguistics and International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 91–99, 2009.

P. Liang, M. I. Jordan, and D. Klein. Learning dependency-based compositional semantics. In *Association for Computational Linguistics (ACL)*, pages 590–599, 2011.

Percy Liang and Cristopher Potts. Bringing machine learning and compositional semantics together. *Annual Review of Linguistics*, 1(1):355–376, 2015.

Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03)*, 2003.

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

Changsong Liu and Joyce Yue Chai. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2288–2294, 2015.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. Discriminating image senses by clustering with multimodal features. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 547–554, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. Automated two-way entrainment to improve spoken dialog system performance. In *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2013.

Dongcai Lu and Xiaoping Chen. Towards an architecture combining grounding and planning for human-robot interaction. In *RoboCup*, pages 214–225, 2015.

Nichola Lubold, Erin Walker, and Heather Pon-Barry. Relating entrainment, grounding, and topic of discussion in collaborative learning dialogues. In *Proceedings of the 11th International Conference on Computer Supported Collaborative Learning (CSCL)*, 2015.

Aurelien Lucchi and Jason Weston. Joint image and word sense discrimination for image retrieval. In *Computer Vision–ECCV*, pages 130–143. Springer, 2012.

Suresh Manandhar, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan. Semeval-2010 task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pages 63–68, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012a.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *International Symposium on Experimental Robotics (ISER)*, 2012b.

Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, Quebéc City, Quebéc, Canada, 2014.

Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. Listen, attend, and walk: Neural mapping of navigational instructions to action sequences. In *Proceedings of AAAI*, 2016.

Çetin Meriçli, Steven D. Klee, Jack Paparian, and Manuela Veloso. An interactive approach for situated task specification through verbal instructions. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, pages 1069–1076, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-2738-1.

Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to mobile manipulation instructions. In *Robotics: Science and Systems*, RSS, 2014.

Shiwali Mohan, Aaron H Mininger, James R Kirk, and John E Laird. Acquiring grounded representations of words with situated interactive instruction. In *Advances in Cognitive Systems*, 2012.

Shiwali Mohan, Aaron H. Mininger, and John E. Laird. 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, 2013.

Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2): 10, 2009.

Aishwarya Padmakumar, Jesse Thomason, and Raymond J. Mooney. Integrated learning of dialog strategies and semantic parsing. 2016.

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

Ted Pedersen and Rebecca Bruce. Distinguishing word senses in untagged text. In *Proceedings of the Second Conference on Empirical Methods in Natural Language*, pages 197–207, 1997.

Ian Perera and James F. Allen. Sall-e: Situated agent for language learning. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, pages 1241–1247, Bellevue, Washington, USA, 2013.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source robot operating system. In *Open Source Software in Robotics Workshop at the IEEE International Conference on Robotics and Automation (ICRA)*, 2009.

Joseph Reisinger and Raymond J. Mooney. Multi-prototype vector-space models of word meaning. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-10)*, pages 109–117, 2010.

Rodolfo Rodriquez, Jesse Thomason, and Raymond J. Mooney. Using semantic parsing to aid in speech recognition. 2016.

Andrew Rosenberg and Julia Hirschberg. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, 2007.

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

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

Kate Saenko and Trevor Darrell. Unsupervised learning of visual sense models for polysemous words. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1393–1400. 2008.

Hinrich Schutze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

Lanbo She, Shaohua Yang, Yu Cheng, Yunyi Jia, Joyce Y. Chai, and Ning Xi. Back to the blocks world: Learning new actions through situated human-robot dialogue. In *Proceedings of 15th SIGDIAL Meeting on Discourse and Dialogue*, 2014.

Carina Silberer and Mirella Lapata. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Jivko Sinapov, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 62(5):632–645, 2014a.

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

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

Mark Steedman and Jason Baldridge. Combinatory categorial grammar. In Robert Borsley and Kersti Borjars, editors, *Non-Transformational Syntax: Formal and Explicit Models of Grammar*. Wiley-Blackwell, 2011.

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

Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. *Proceedings of Robotics: Science and Systems*, 2014.

Jesse Thomason and Raymond Mooney. Multi-modal word synset induction. 2016.

Jesse Thomason, Huy Nguyen, and Diane Litman. Prosodic entrainment and tutoring dialogue success. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED)*, pages 750–753, July 2013.

Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1923–1929, July 2015.

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

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

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

Matthew Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. Learning semantic maps from natural language descriptions. *Proceedings of Robotics: Science and Systems*, 2013.

David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. Interpreting Multimodal Referring Expressions in Real Time. In *International Conference on Robotics and Automation*, 2016.

Yezhou Yang, Cornelia Fermüller, Yiannis Aloimonos, and Anupam Guha. A cognitive system for understanding human manipulation actions. *Advances in Cognitive Systems*, 3:67–86, 2014.

David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL95*, pages 189–196, 1995.

Steve Young, Milica Gasic, Blaise Thomson, and Jason D Williams. POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.