Continuously Improving Robotic Natural Language Understanding with Semantic Parsing, Dialog, and Multi-modal Perception

Jesse Thomason
University of Texas at Austin

Natural Language Understanding for Robots

- Robots are increasingly present in human environments
  - Stores, hospitals, factories, and offices
- People communicate in natural language
- Robots should understand natural language commands from humans

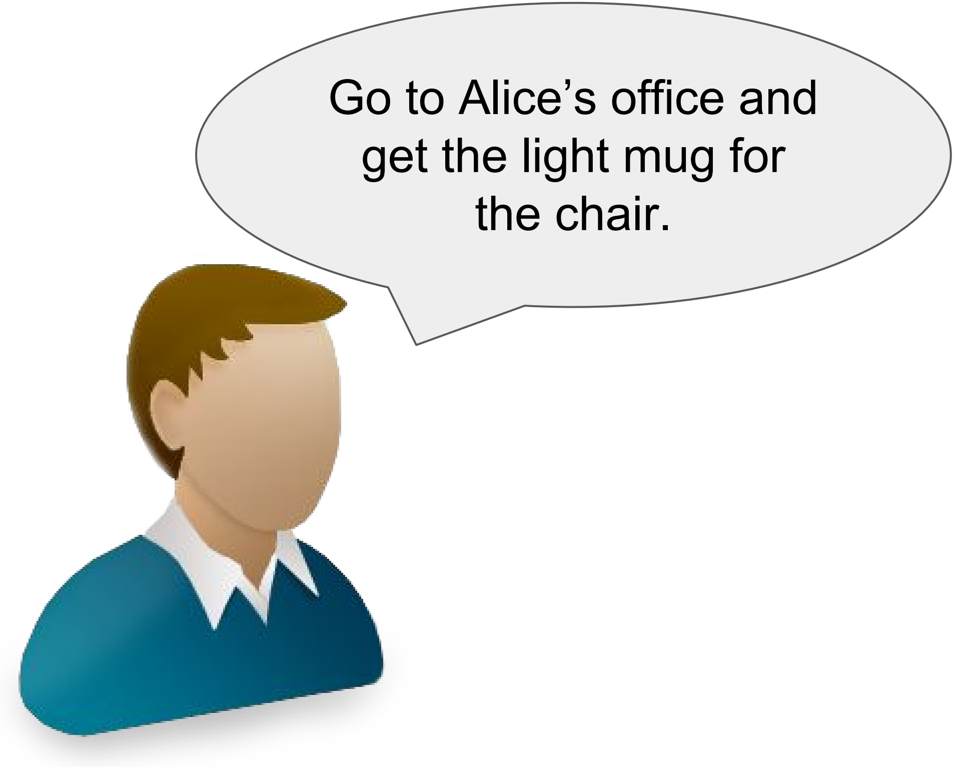
Natural Language Understanding for Robots

- Different robots have different sensing and manipulation capabilities
- Different domains require understanding different vocabularies
- Learning paradigms can be applicable across platforms and domains

“alert me if her heart rate decreases”
“bring me his chart”
“go and get the family”
“scalpel”

“text me when the speaker arrives”
“grab the heavy, green mug”
“lead him to alice’s office”
“get out of the way”

Natural Language Understanding for Robots



Natural Language Understanding for Robots

Go to Alice’s office and get the light mug for the chair.

- Commands that need to be actualized through robot action

Natural Language Understanding for Robots

Go to Alice’s office and
get the light mug for
the chair.

- Commands that need to be actualized through robot action
- World knowledge about people and the surrounding office space

Natural Language Understanding for Robots

> Go to Alice’s office and
> get the light mug for
> the chair.

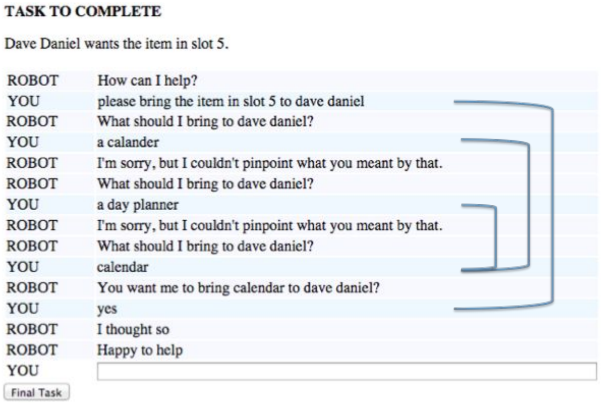
- Commands that need to be actualized through robot action
- World knowledge about people and the surrounding office space
- Perception information to identify referent object

Natural Language Understanding for Robots

- As much as possible, solve these problems independent of robot and domain
- Interaction with humans can strengthen understanding over time
- Use human-robot dialog and interaction as a learning signal

Outline

Learning to Interpret Natural Language Commands through Human-Robot Dialog [Thomason et al. IJCAI 2015]

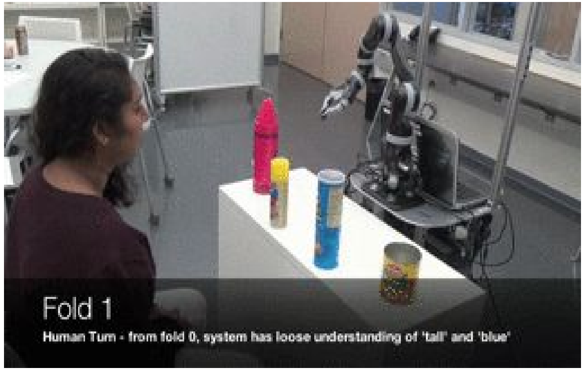


TASK TO COMPLETE

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	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	

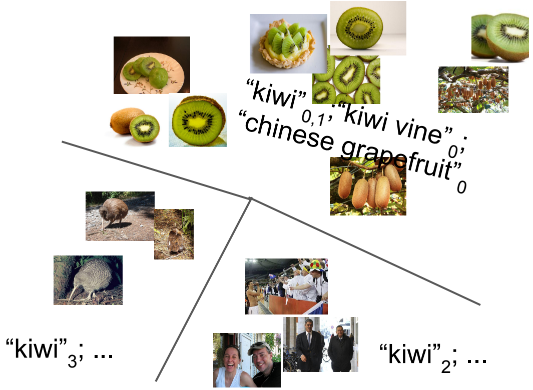
Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy" [Thomason et al. IJCAI 2016]



Fold 1

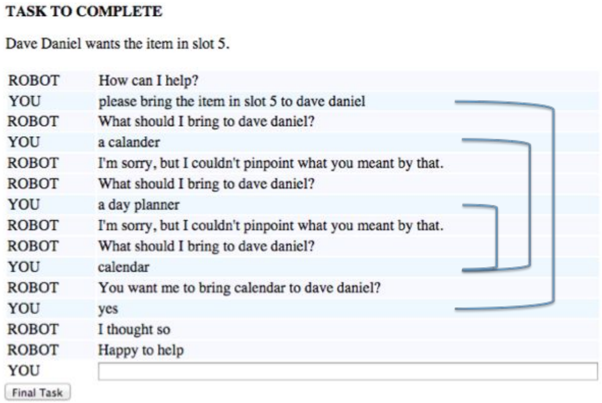
Human Turn - from fold 0, system has loose understanding of 'tall' and 'blue'

Multi-Modal Word Synset Induction [Thomason, Mooney IJCAI 2017]



Outline

Learning to Interpret Natural Language Commands through Human-Robot Dialog [Thomason et al. IJCAI 2015]



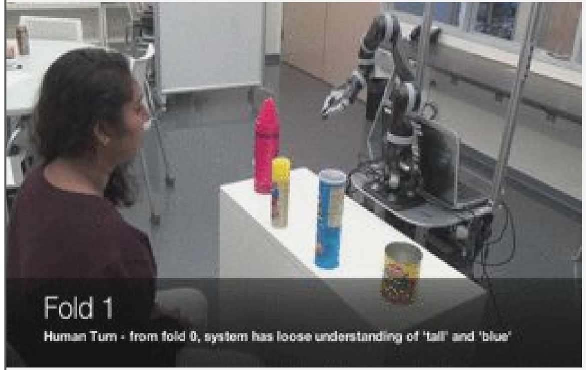
TASK TO COMPLETE

Dave Daniel wants the item in slot 5.

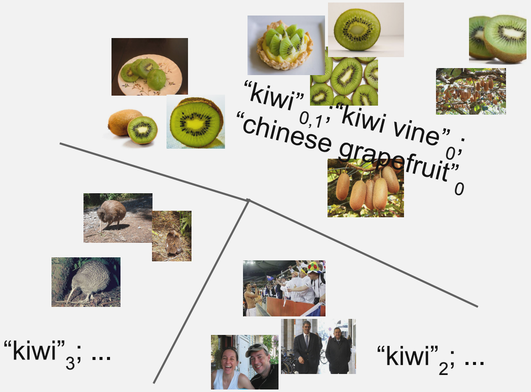
ROBOT	How can I help?
YOU	please bring the item in slot 5 to dave daniel
ROBOT	What should I bring to dave daniel?
YOU	a calander
ROBOT	I'm sorry, but I couldn't pinpoint what you meant by that.
ROBOT	What should I bring to dave daniel?
YOU	a day planner
ROBOT	I'm sorry, but I couldn't pinpoint what 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

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy" [Thomason et al. IJCAI 2016]


Fold 1
Human Turn - from fold 0, system has loose understanding of 'tall' and 'blue'

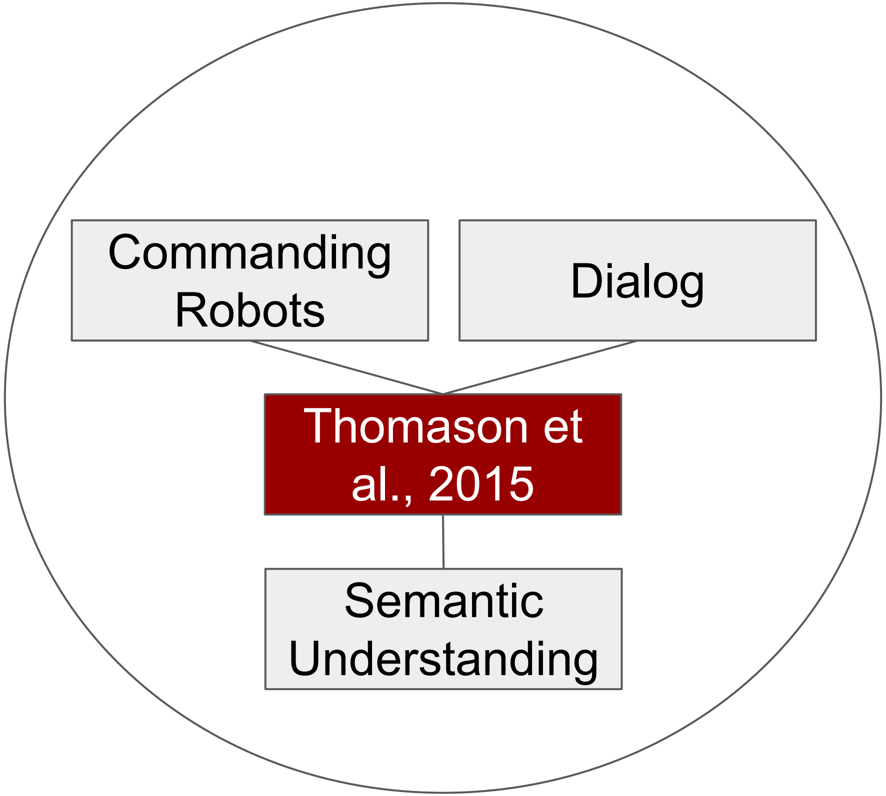
Multi-Modal Word Synset Induction [Thomason, Mooney IJCAI 2017]

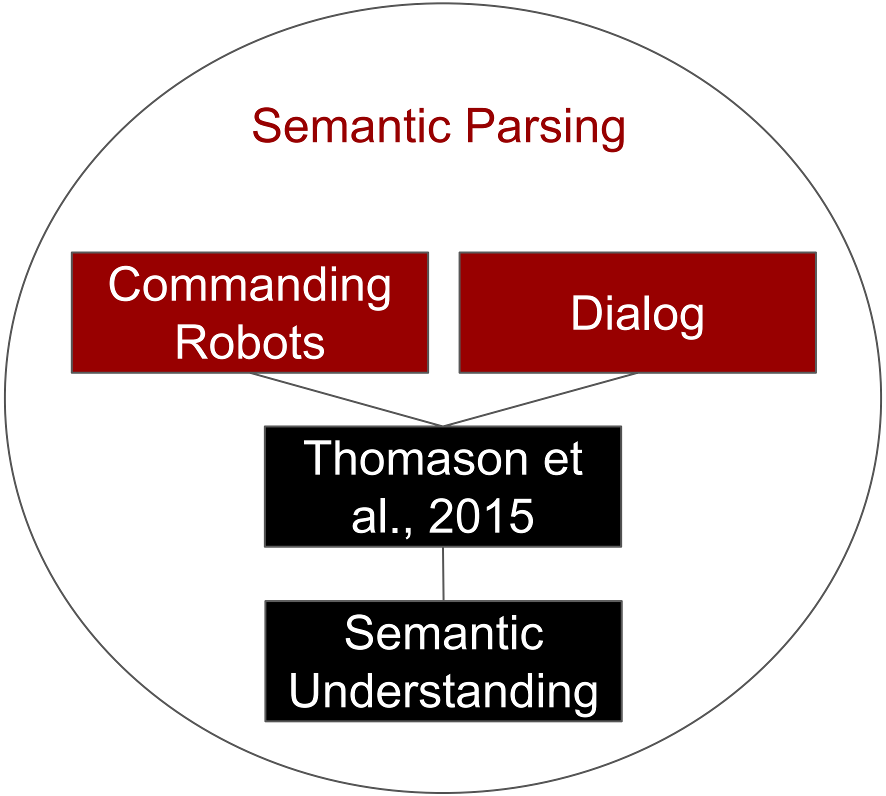


Learning to Interpret Natural Language Commands through Human-Robot Dialog

- Initialize language understanding with minimal resources
- Use human-robot dialogs to get text/semantic form pairs
- Improve semantic parsing over time by retraining on induced pairs

Learning to Interpret Natural Language Commands through Human-Robot Dialog





Semantic Parsing

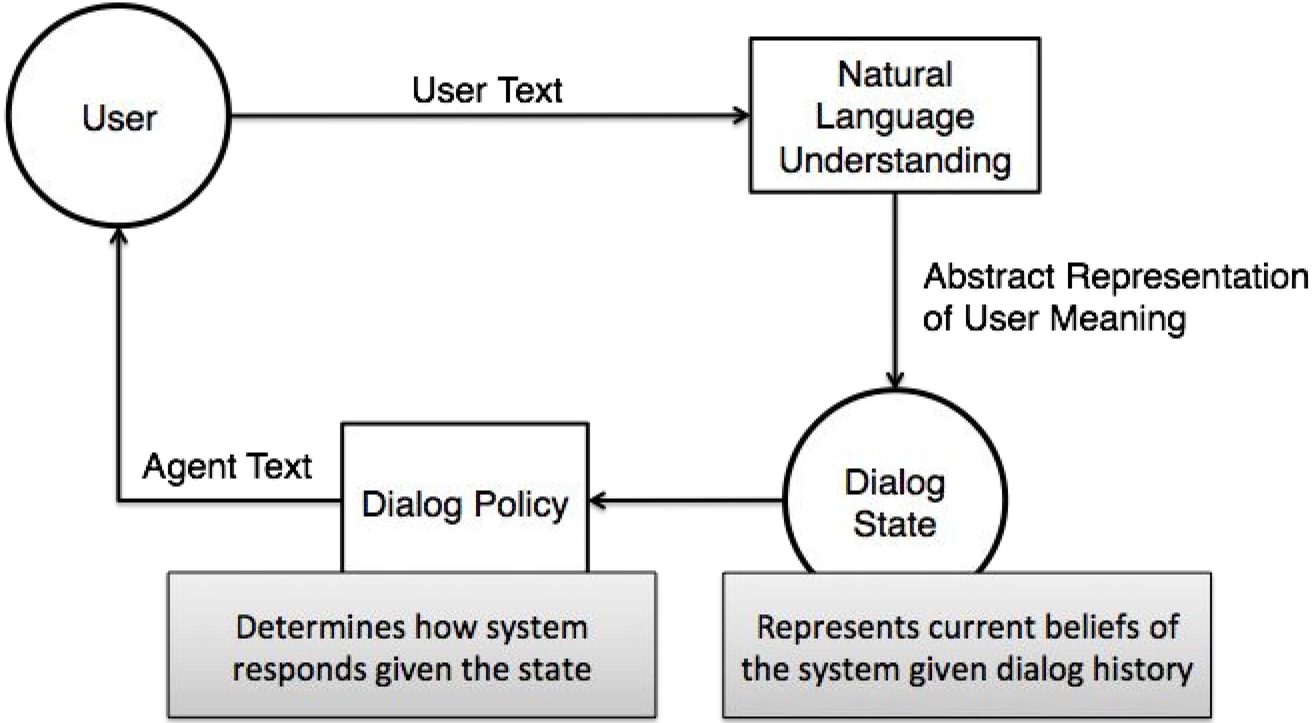
Commanding Robots

Dialog

Thomason et al., 2015

Semantic Understanding

Dialog

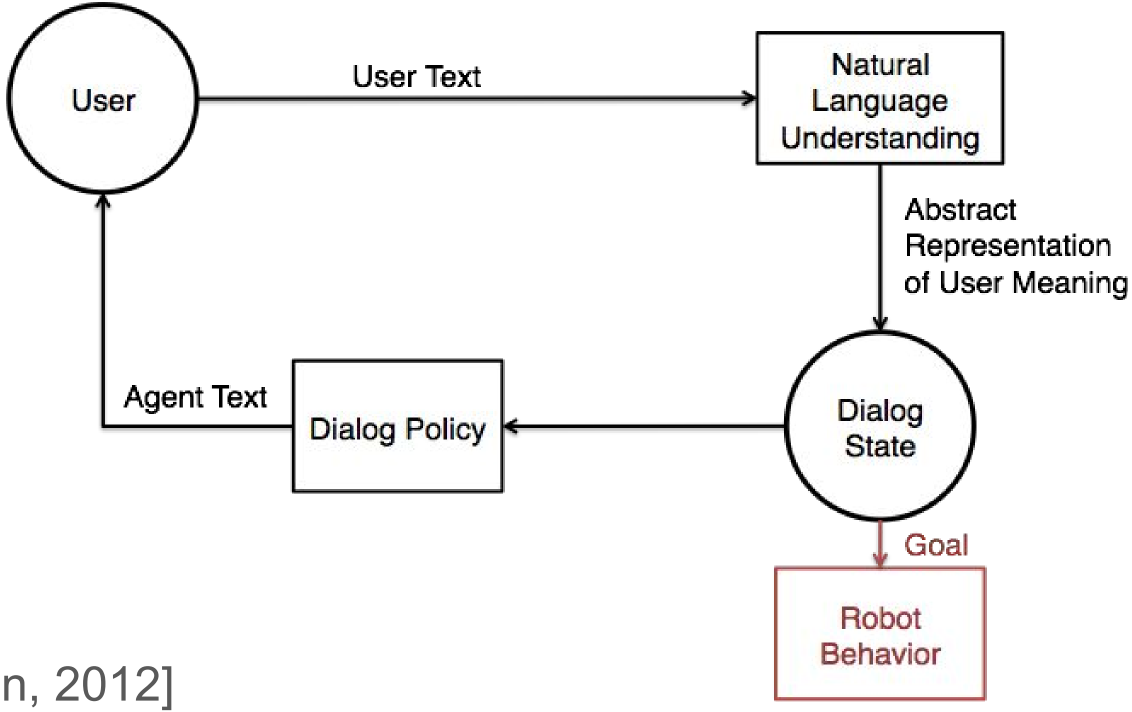


User
User Text
Natural Language Understanding
Abstract Representation of User Meaning
Dialog State
Dialog Policy
Agent Text

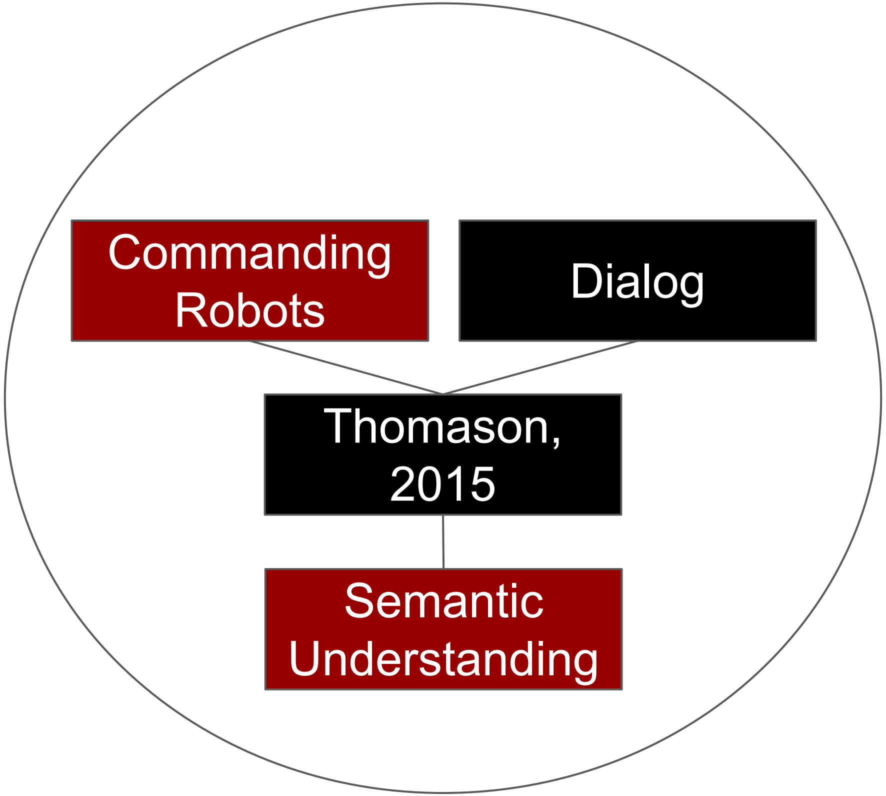
Determines how system responds given the state

Represents current beliefs of the system given dialog history

+ Commanding Robots


User
User Text
Natural Language Understanding
Abstract Representation of User Meaning
Dialog State
Dialog Policy
Agent Text
Goal
Robot Behavior

Dialog can be used for commanding robots [Matuszek, 2012; Mohan, 2012]



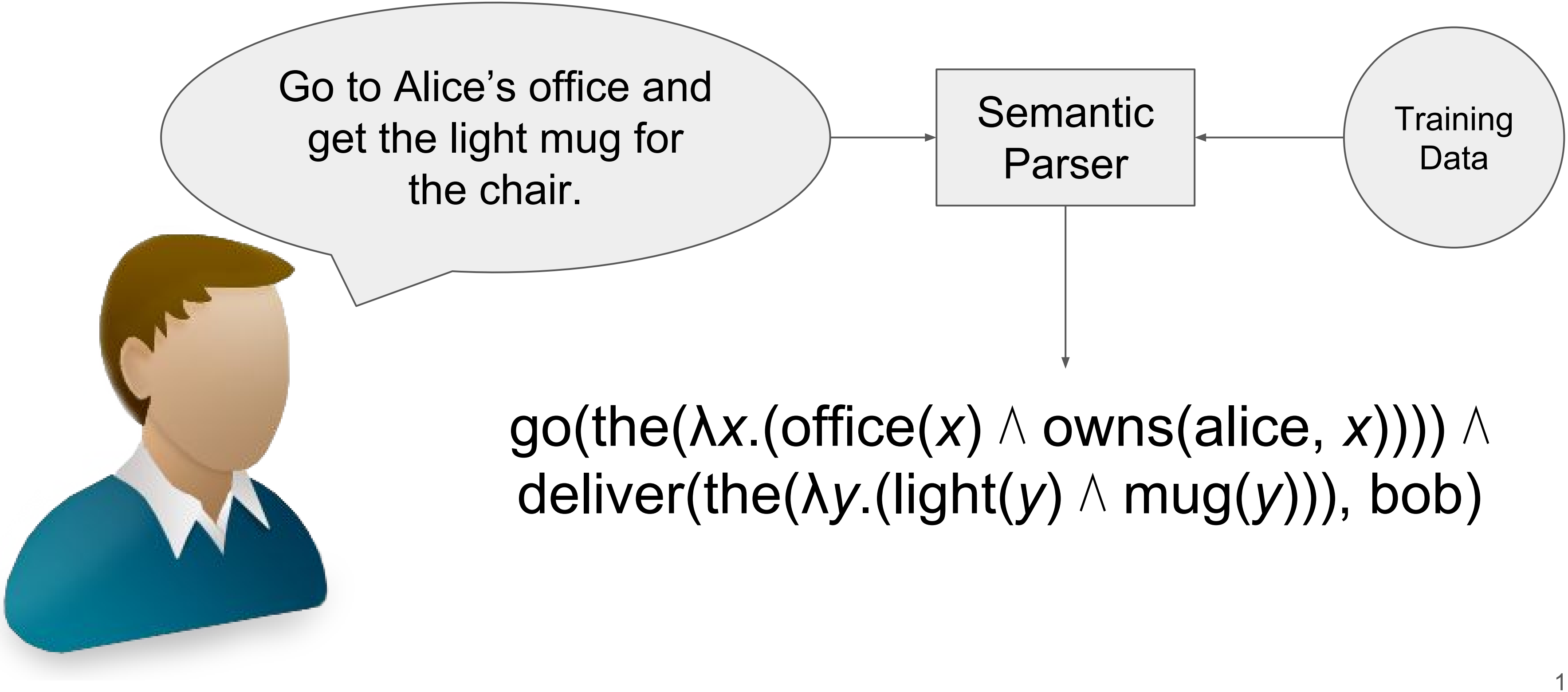
Commanding Robots

Dialog

Thomason, 2015

Semantic Understanding

Background: Semantic Parsing



Go to Alice’s office and get the light mug for the chair.

Semantic Parser

Training Data

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

Background: Semantic Parsing

- Translate from human language to formal language
- We use combinatory categorial grammar formalism [Zettlemoyer, 2005]
- Words assigned part-of-speech-like categories
- Categories combine to form syntax of utterance

Background: Semantic Parsing

- Small example of composition

Alice ‘s office

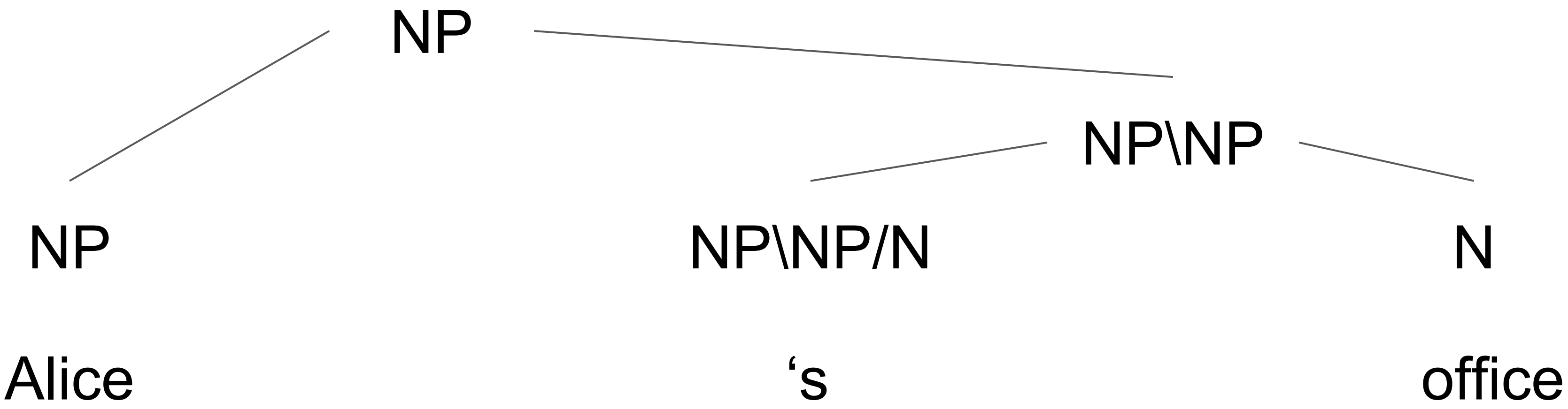
Background: Semantic Parsing

- Small example of composition
- Add part-of-speech-like categories

NP	NP\NP/N	N
Alice	‘s	office

Background: Semantic Parsing

- Add part-of-speech-like categories
- Categories combine right (/) and left (\) to form trees



NP

NP\NP

NP

NP\NP/N

N

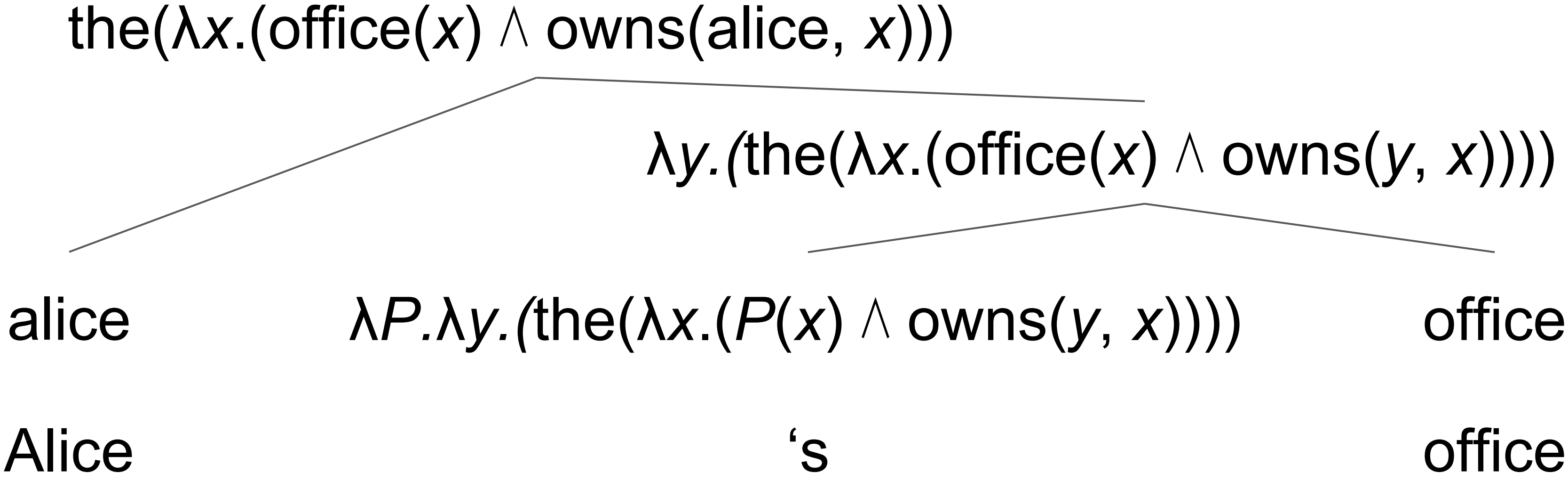
Alice

‘s

office

Background: Semantic Parsing

- Leaf-level semantic meanings can be propagated through tree



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

$\lambda y.(\text{the}(\lambda x.(\text{office}(x) \land \text{owns}(y, x))))$

alice $\lambda P.\lambda y.(\text{the}(\lambda x.(P(x) \land \text{owns}(y, x))))$ office

Alice ‘s office

Background: Semantic Parsing

- `get’ refers to the action predicate deliver
- `light’ could mean light in color or light in weight
- bob is referred to as `the chair’, his title

Go to Alice’s office and
get the light mug for
the chair.

$$\mathrm{go}(\mathrm{the}(\lambda x.(\mathrm{office}(x) \wedge \mathrm{owns}(\mathrm{alice}, x)))) \wedge$$
$$\mathrm{deliver}(\mathrm{the}(\lambda y.(\mathrm{light2}(y) \wedge \mathrm{mug1\_cup2}(y))), \mathrm{bob})$$

Background: Semantic Parsing

- Parsers can be trained from paired examples
- Sentences and their semantic forms
- Treat underlying tree structure as latent during inference [Liang 2015]
- With pairs of human commands and semantic forms, can train a semantic parser for robots

Background: Semantic Parsing

- Parsers can be trained from paired examples
- For example, parameterize parse decisions in a weighted perceptron model
  - Word -> CCG assignment counts (e.g. “for -> PP/NP; “alice -> NP”)
  - CCG production counts (e.g. “PP -> PP/NP NP”; “S -> NP S\NP)
  - Word -> semantics counts

    (e.g. “the chair -> bob”; “the chair -> the(λ*x*.(chair(x)))”)
- Guide search for best parse using perceptron
- Update parameters during training by contrasting best scoring parse to known true parse; for example using hinge loss

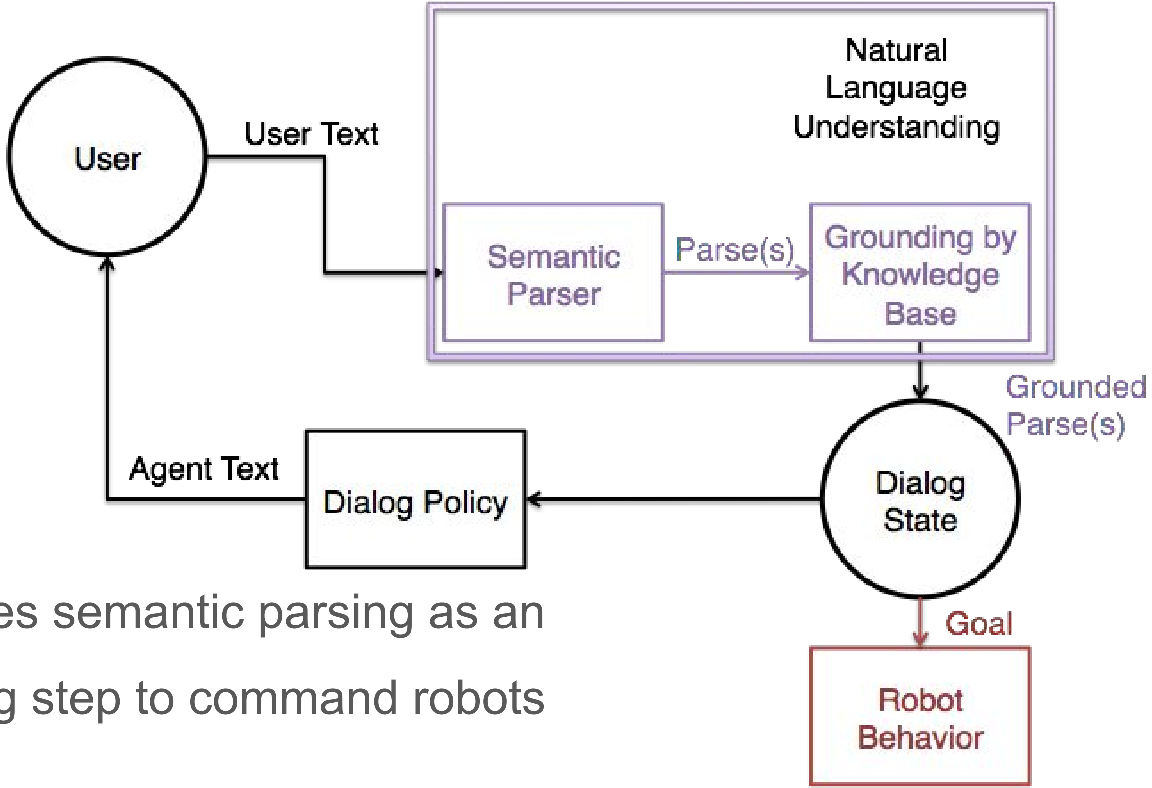
Background: Language Grounding

- Some *x* that is an office and is owned by Alice
- Membership and ownership relations can be kept in a knowledge base
- Created by human annotators to describe surrounding environment

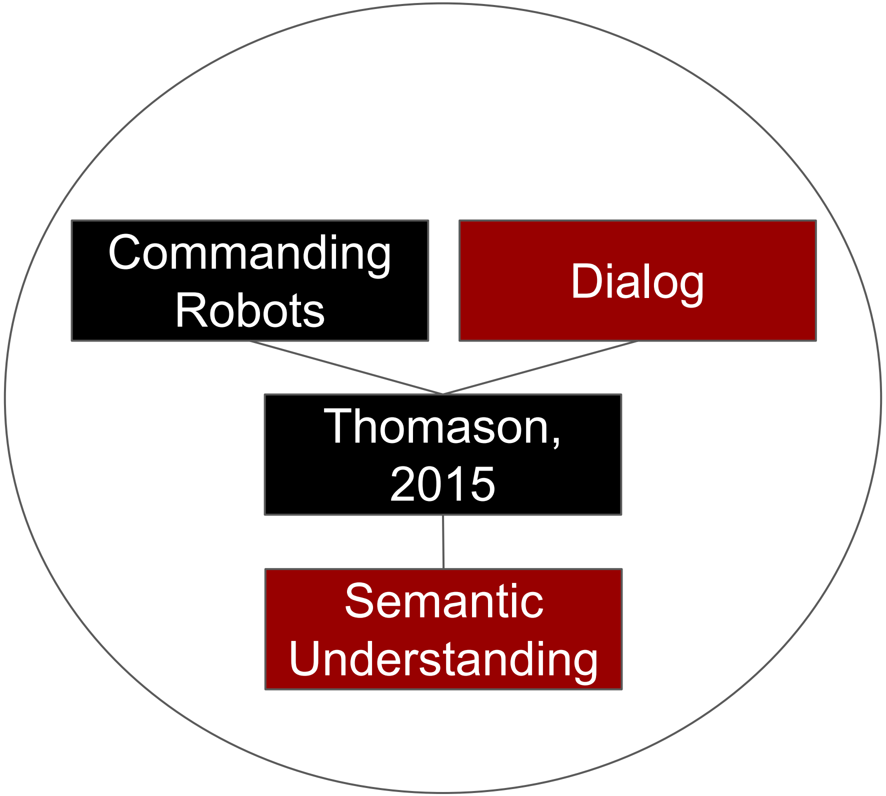
“Alice’s office”

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

+ Semantic Parsing



Past work uses semantic parsing as an understanding step to command robots [Kollar, 2013]



Commanding Robots

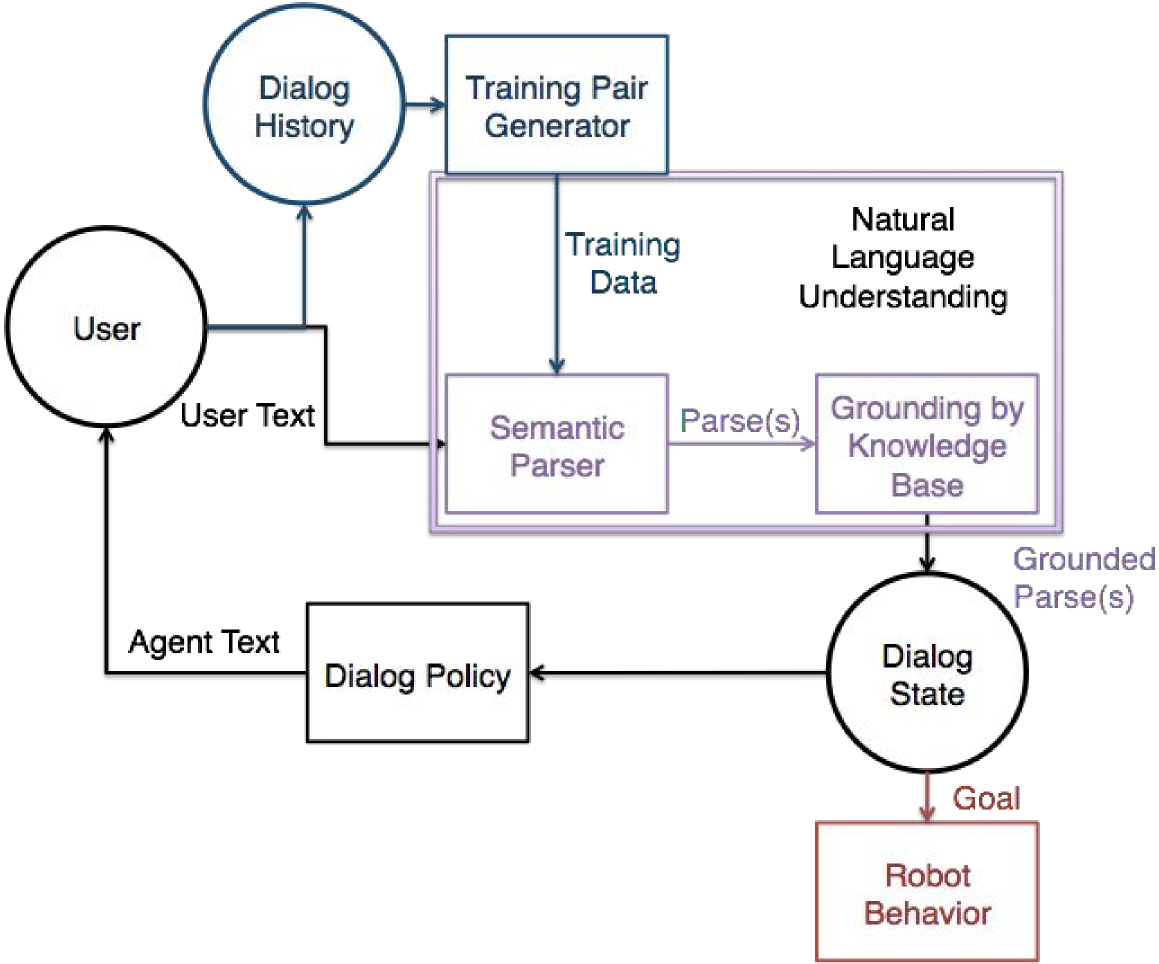
Dialog

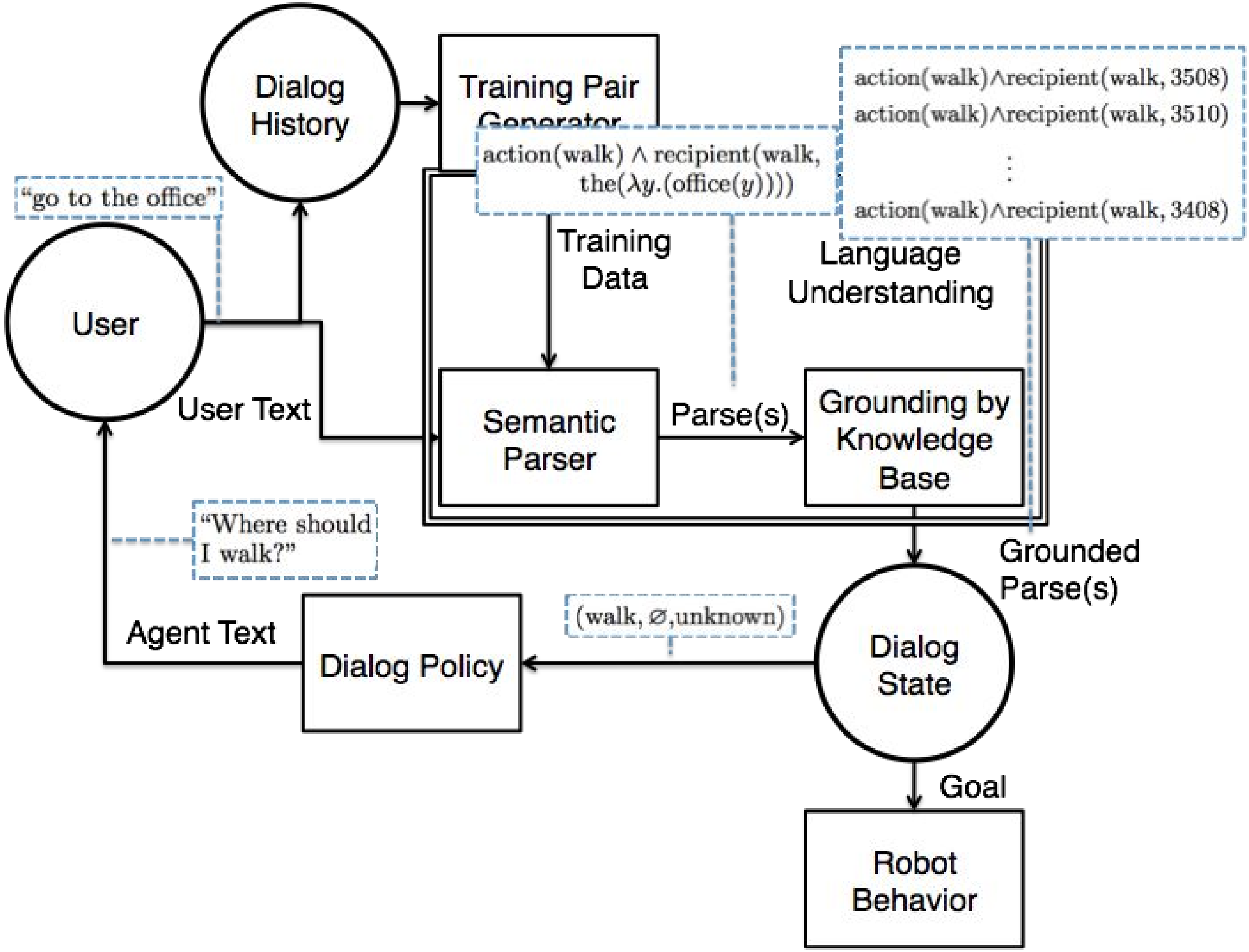
Thomason, 2015

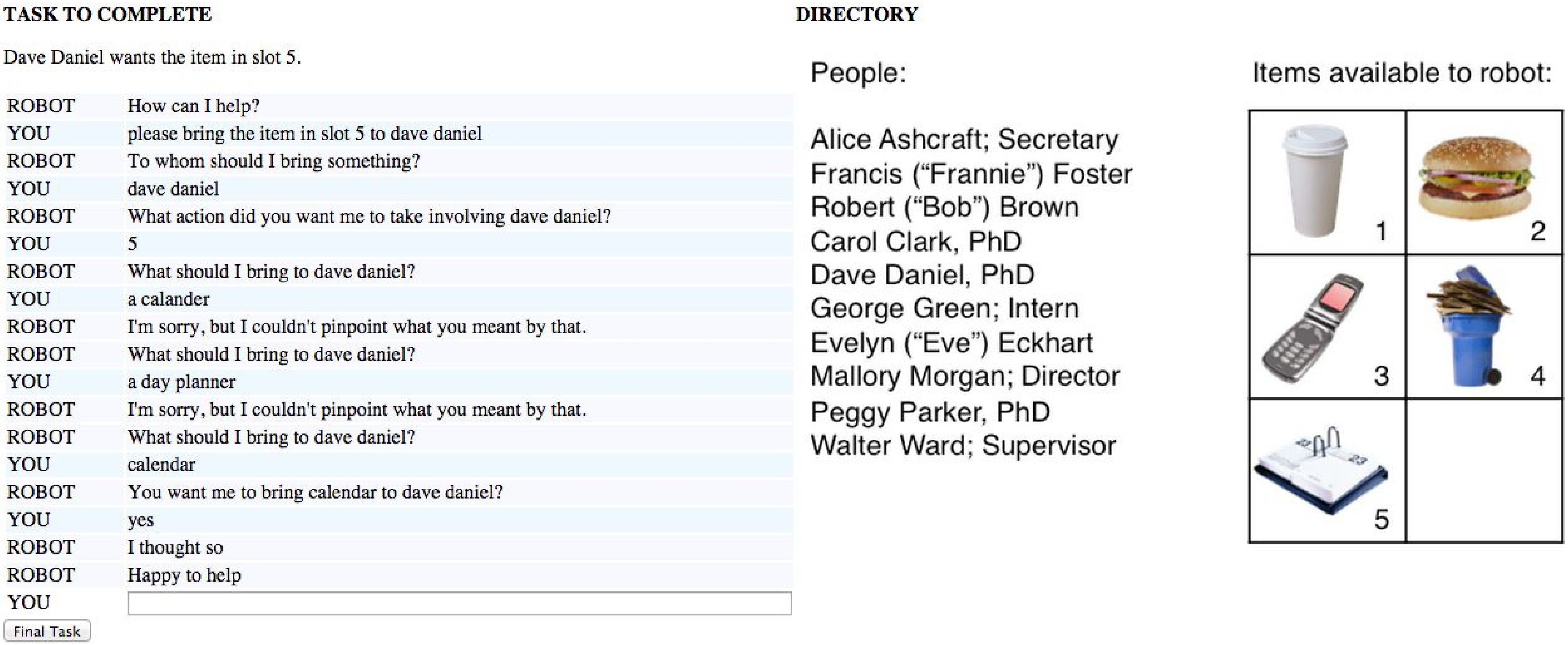
Semantic Understanding

Generating New Training Examples

- Past work generates training data for a parser given a corpus of conversations [Artzi, 2011]
  - In conversations, system utterances are tagged with gold semantic meanings
- We pair confirmed understanding from dialog with previous misunderstandings
  - When system understands meaning Y, assume all past utterances X_i that tried to point to Y meant Y


Dialog History
Training Pair Generator
Training Data
Natural Language Understanding
User
User Text
Semantic Parser
Parse(s)
Grounding by Knowledge Base
Grounded Parse(s)
Agent Text
Dialog Policy
Dialog State
Goal
Robot Behavior


Dialog History
Training Pair Generator
action(walk) ∧ recipient(walk, the(λy.(office(y))))
action(walk)∧recipient(walk, 3508)
action(walk)∧recipient(walk, 3510)
⋮
action(walk)∧recipient(walk, 3408)
"go to the office"
User
Training Data
Language Understanding
User Text
Semantic Parser
Parse(s)
Grounding by Knowledge Base
Grounded Parse(s)
"Where should I walk?"
Agent Text
Dialog Policy
(walk, ∅,unknown)
Dialog State
Goal
Robot Behavior



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	

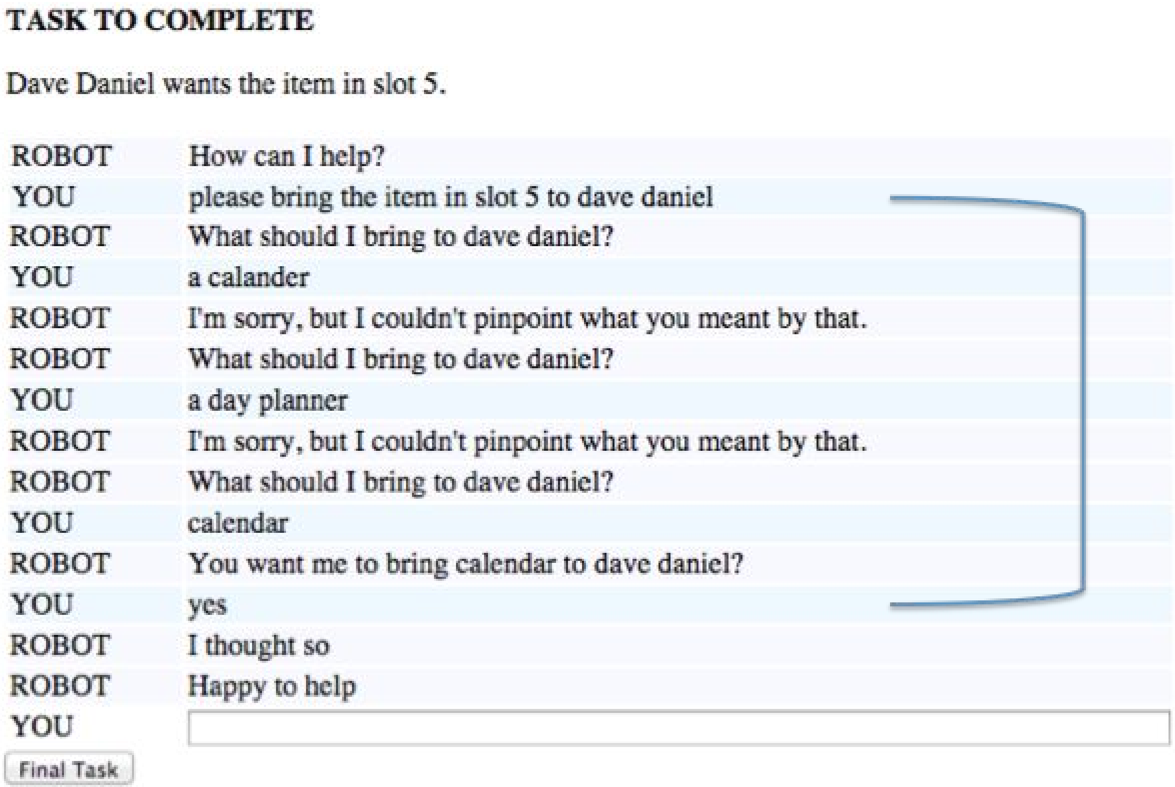
DIRECTORY

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:

Generating New Training Examples

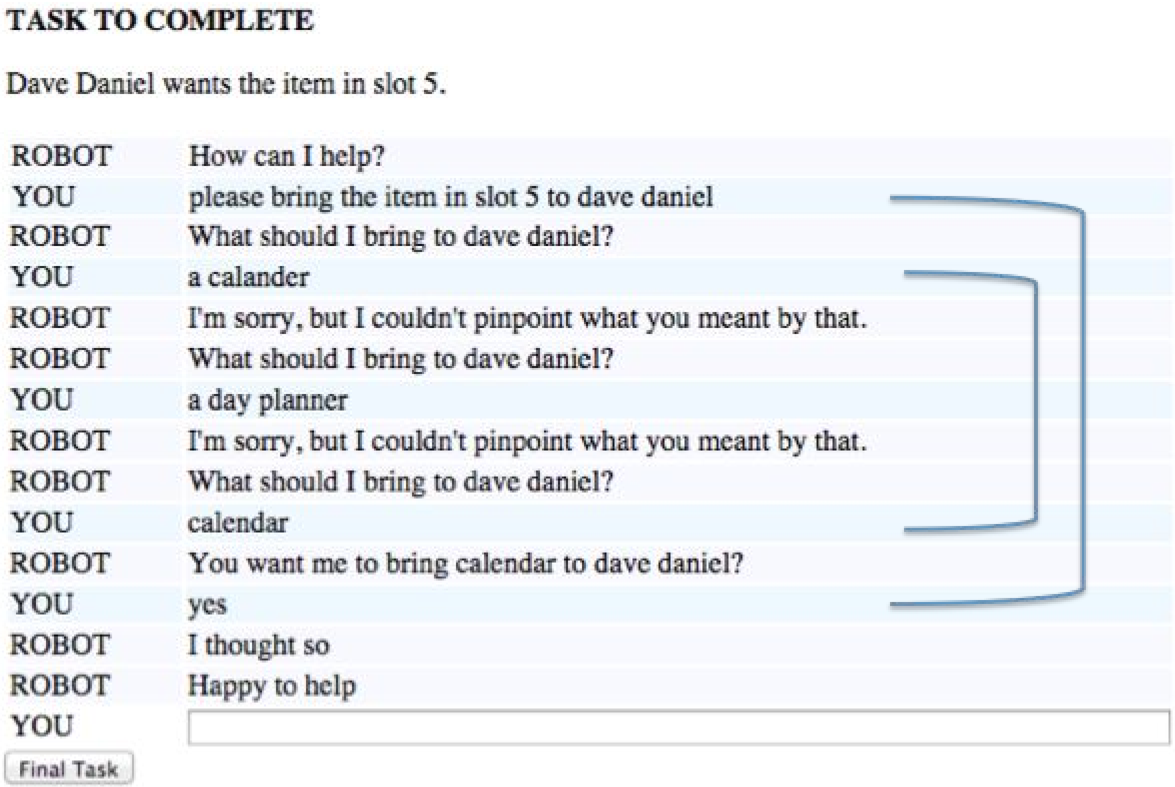


TASK TO COMPLETE

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	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	

Generating New Training Examples

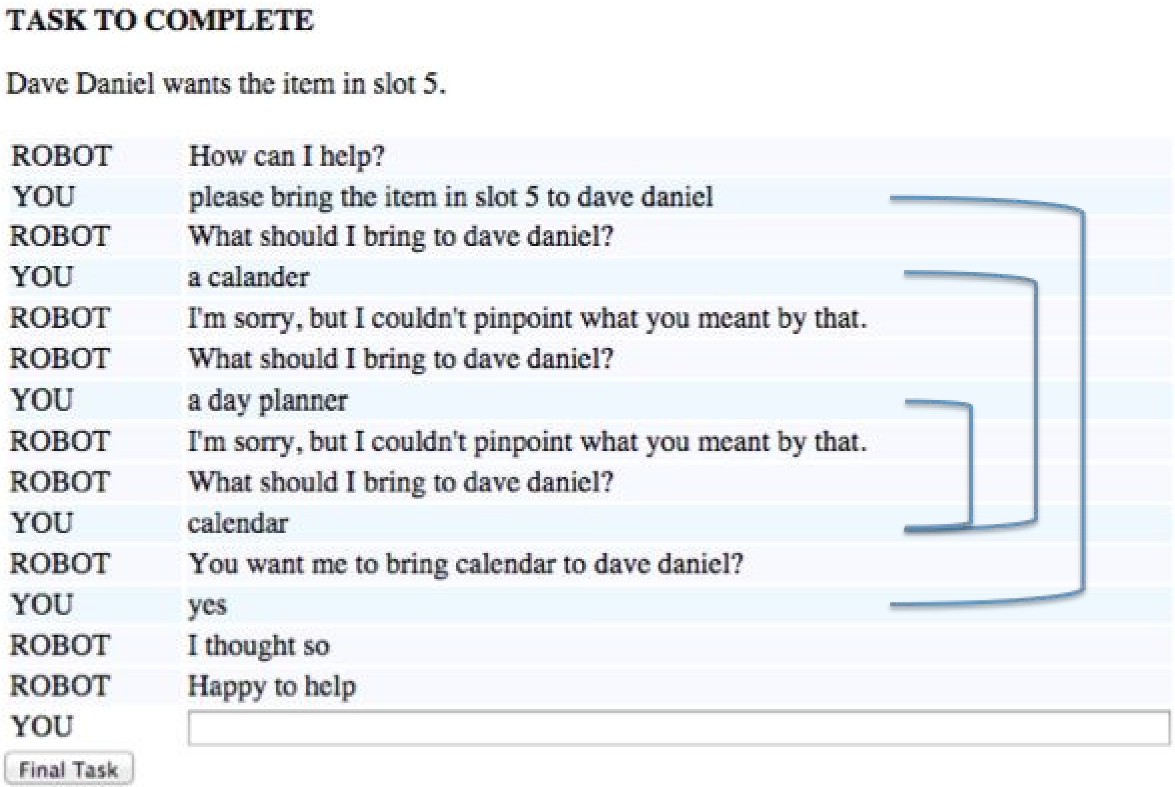


TASK TO COMPLETE

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	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	

Generating New Training Examples

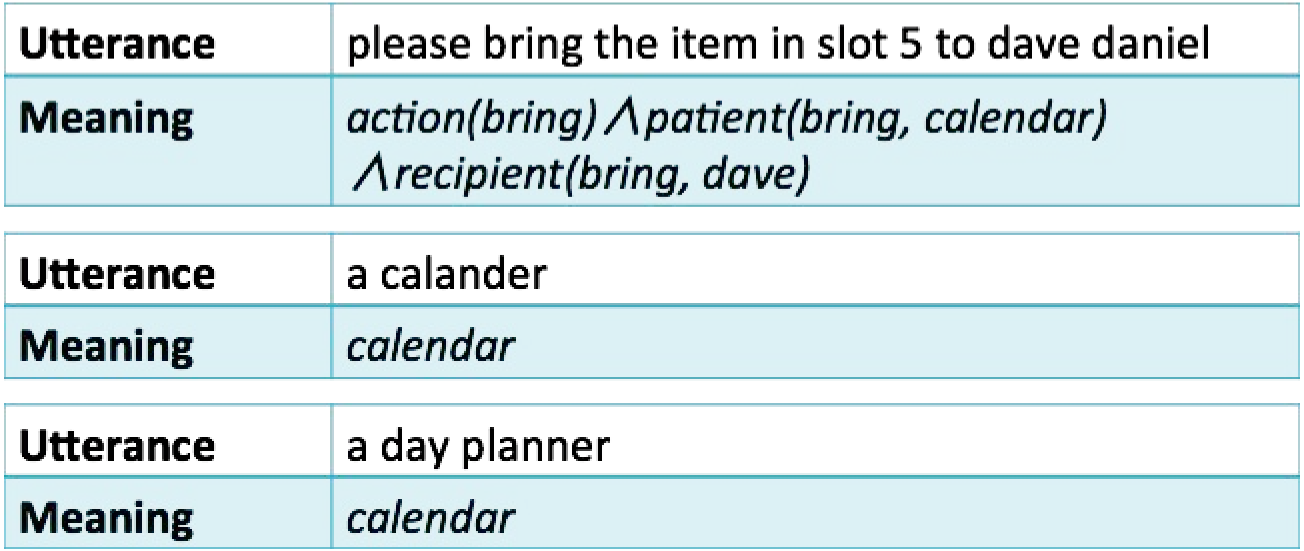


TASK TO COMPLETE

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	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	

Generating New Training Examples

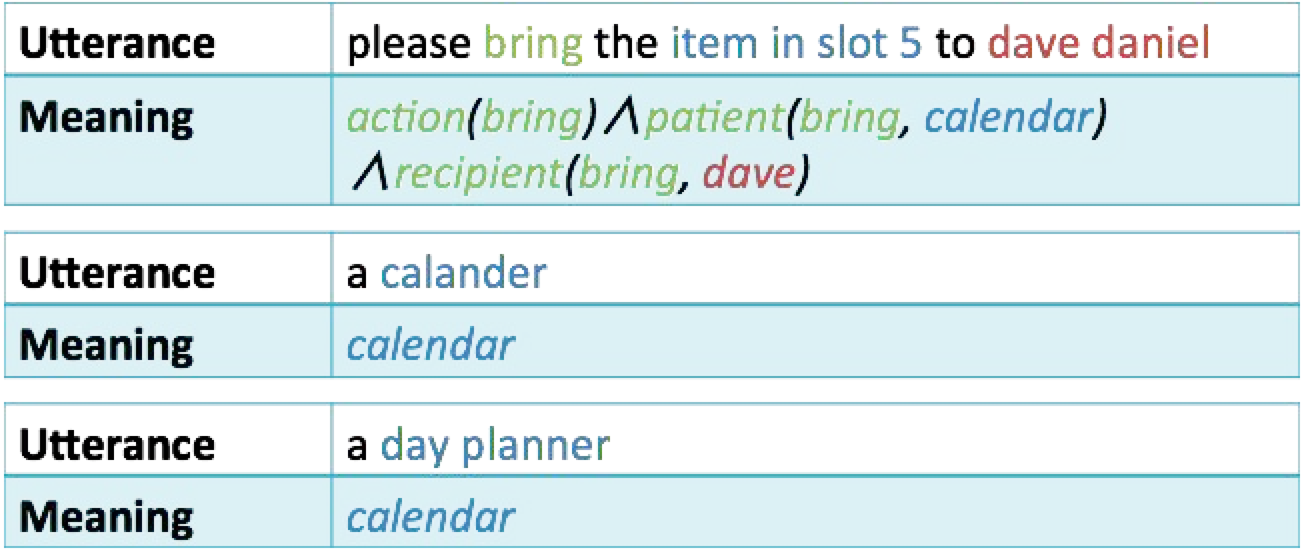


Utterance	please bring the item in slot 5 to dave daniel
Meaning	*action(bring) $\wedge$patient(bring, calendar)* *$\wedge$recipient(bring, dave)*

Utterance	a calander
Meaning	*calendar*

Utterance	a day planner
Meaning	*calendar*

Generating New Training Examples

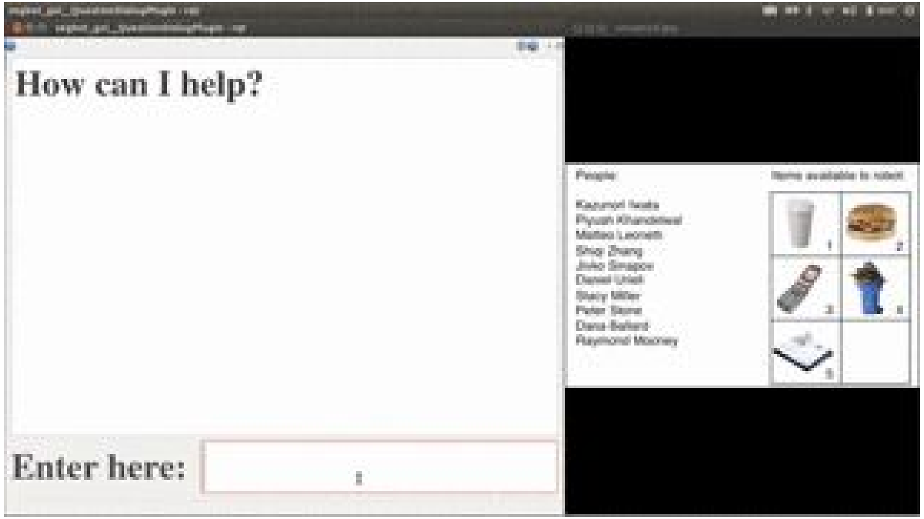


Utterance	please bring the item in slot 5 to dave daniel
Meaning	*action(bring) ∧ patient(bring, calendar) ∧ recipient(bring, dave)*

Utterance	a calander
Meaning	*calendar*

Utterance	a day planner
Meaning	*calendar*

Generating New Training Examples



How can I help?

Enter here:

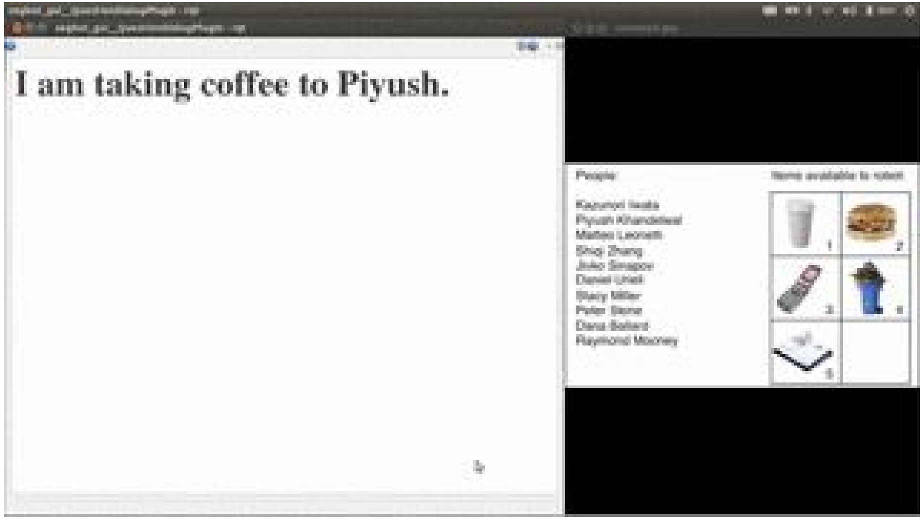
People

Kazunori Iwata
Piyush Khandelwal
Matteo Leonetti
Shiqi Zhang
Jivko Sinapov
Daniel Urieli
Stacy Miller
Peter Stone
Dana Ballard
Raymond Mooney

Items available to robot

Generating New Training Examples

Generating New Training Examples

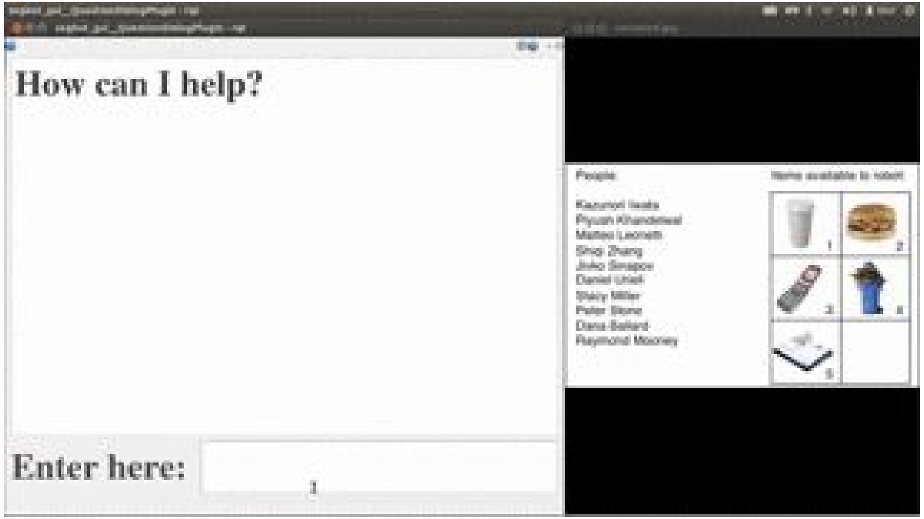


I am taking coffee to Piyush.

People

Kazunori Iwata
Piyush Khandelwal
Matteo Leonetti
Shiqi Zhang
Jivko Sinapov
Daniel Urieli
Stacy Miller
Peter Stone
Dana Ballard
Raymond Mooney

Generating New Training Examples



How can I help?

Enter here:

People

Kazunori Iwata
Piyush Khandelwal
Matteo Leonetti
Shiqi Zhang
Jivko Sinapov
Daniel Urieli
Stacy Miller
Peter Stone
Dana Ballard
Raymond Mooney

Items available to robot:

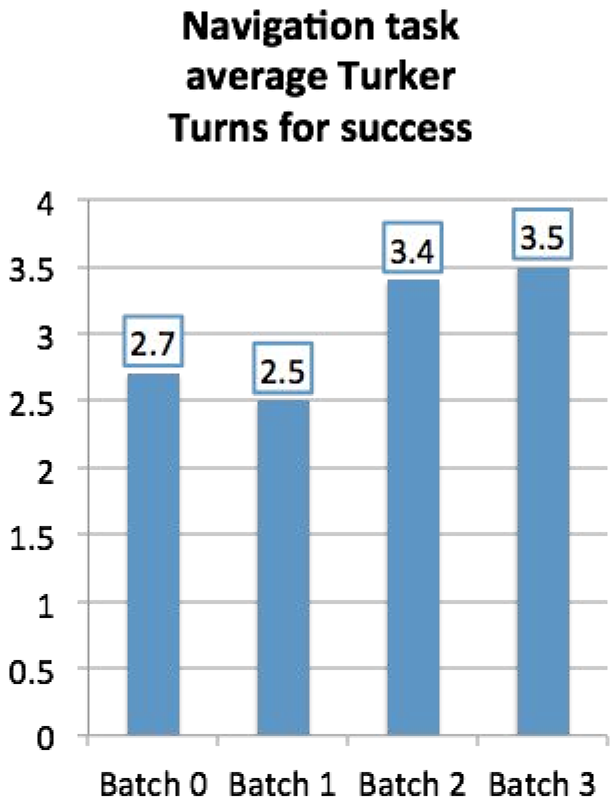
Experiments

- Hypothesis: Performing incremental re-training of a parser with sentence/parse pairs obtained through dialog will result in better user experience than using a pre-trained parser alone
- Tested via:
    - Mechanical Turk - many users, unrealistic interaction (just text, no robot)
    - Segbot Platform - few users, natural interactions with real world robot

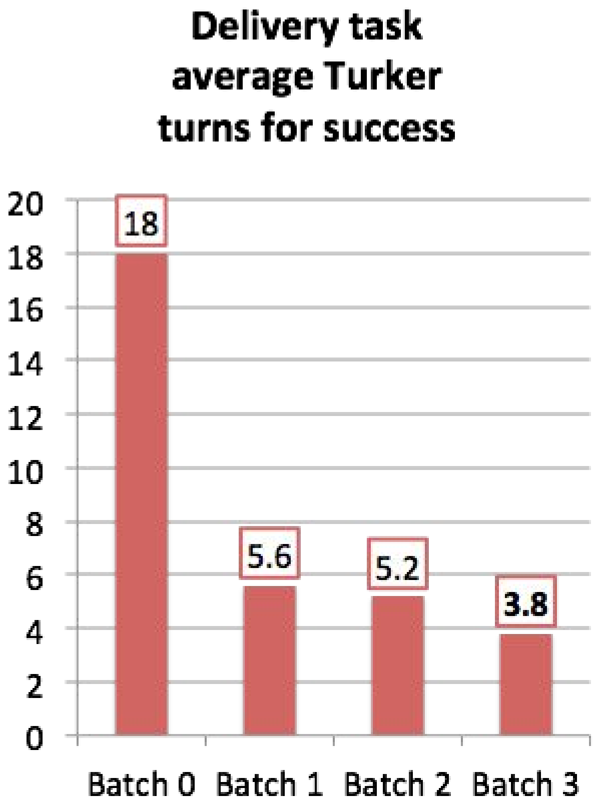
Mechanical Turk Experiment

- Four batches of ~100 users each
- Retraining after every batch (~50 training goals)
- Performance measured every batch (~50 testing goals)
- Goals:
    - Navigation - user told the robot is needed in a certain room (one action, single argument)
    - Delivery - user told a certain person needs a certain item (one action, two arguments)

Mechanical Turk Dialog Turns

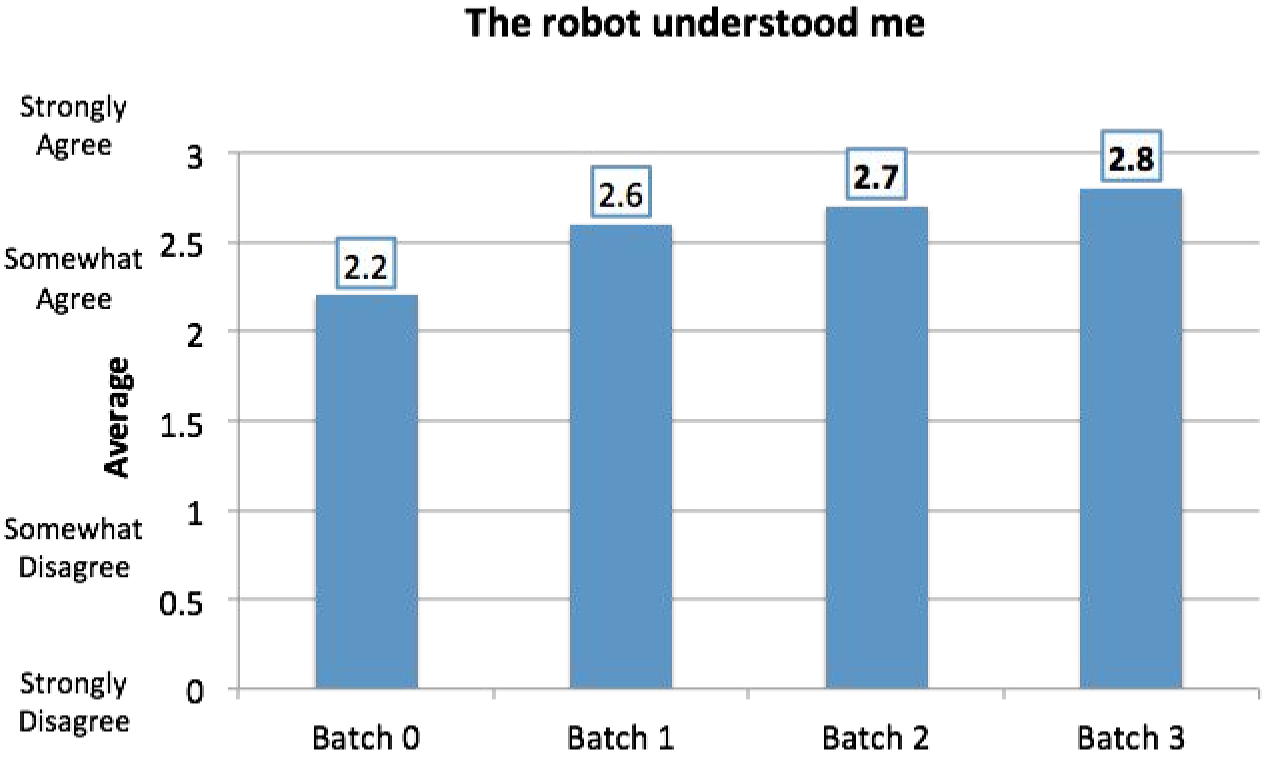


Navigation task average Turker Turns for success

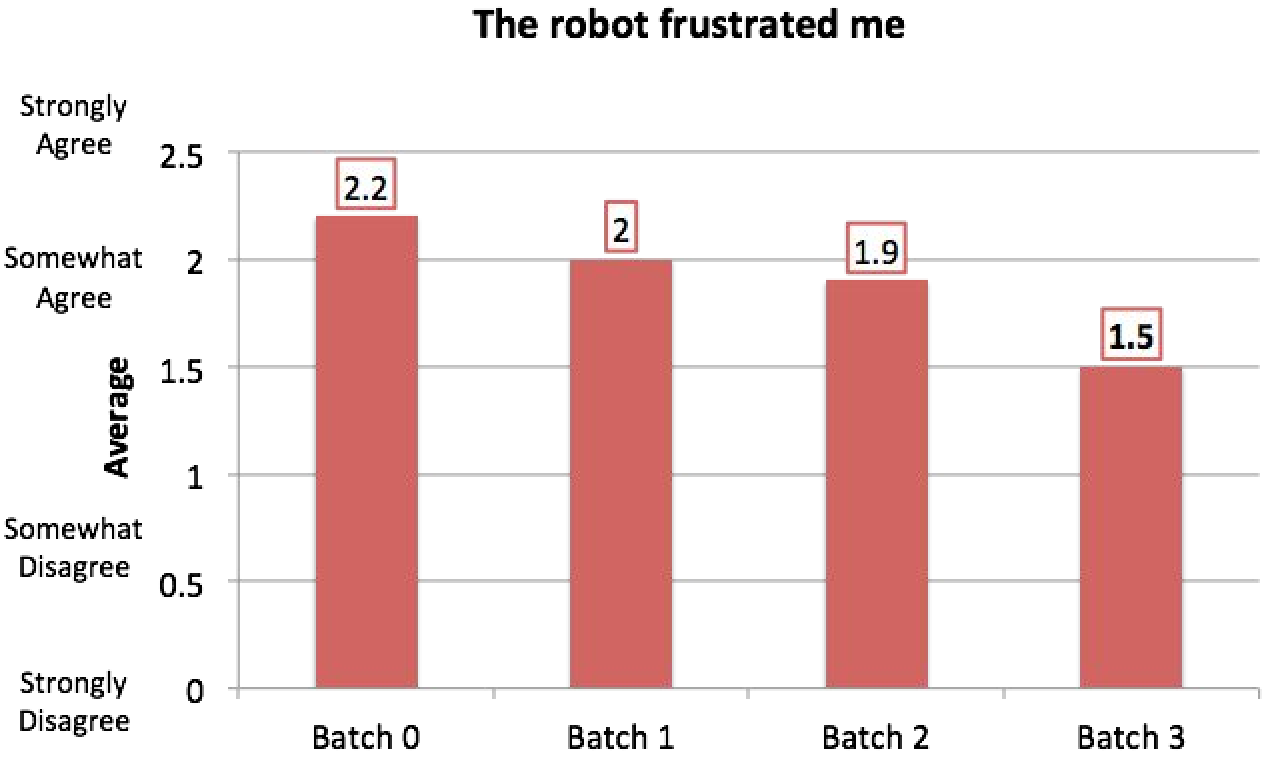


Delivery task average Turker turns for success

Mechanical Turk Survey Responses



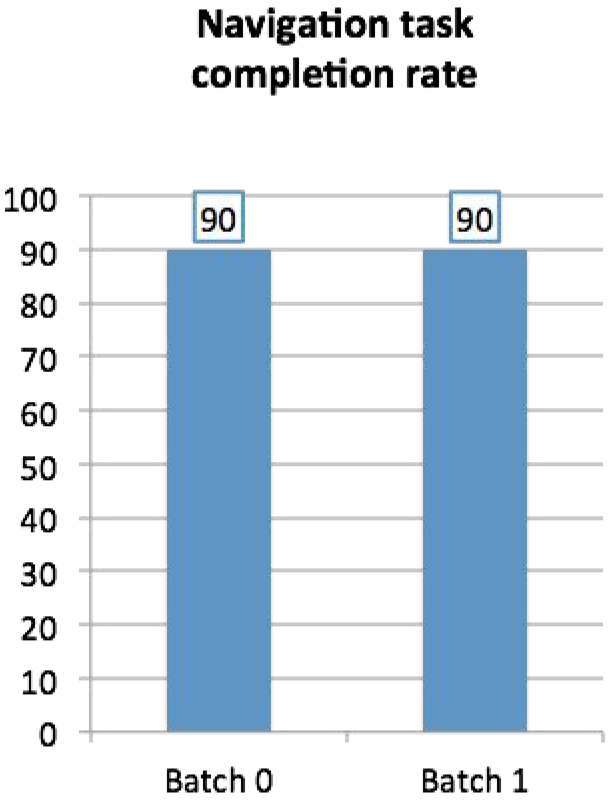
Mechanical Turk Survey Responses

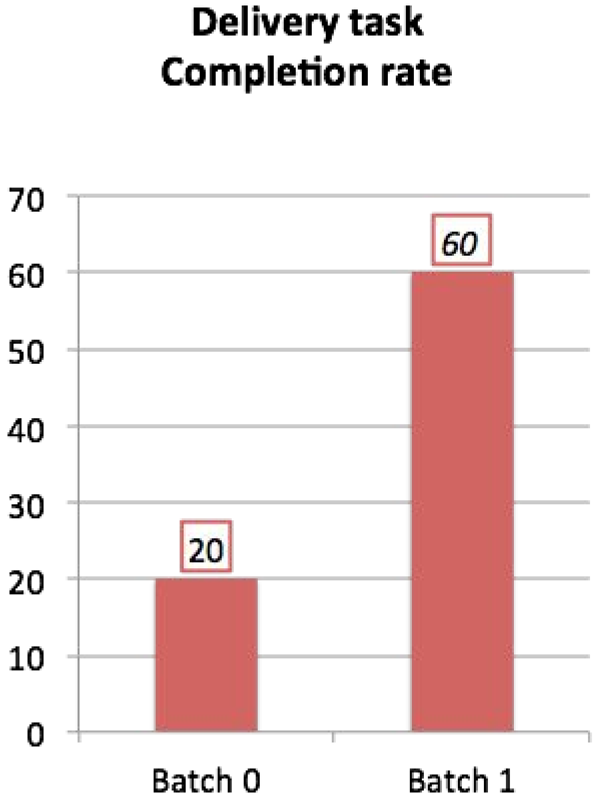


Segbot Experiment

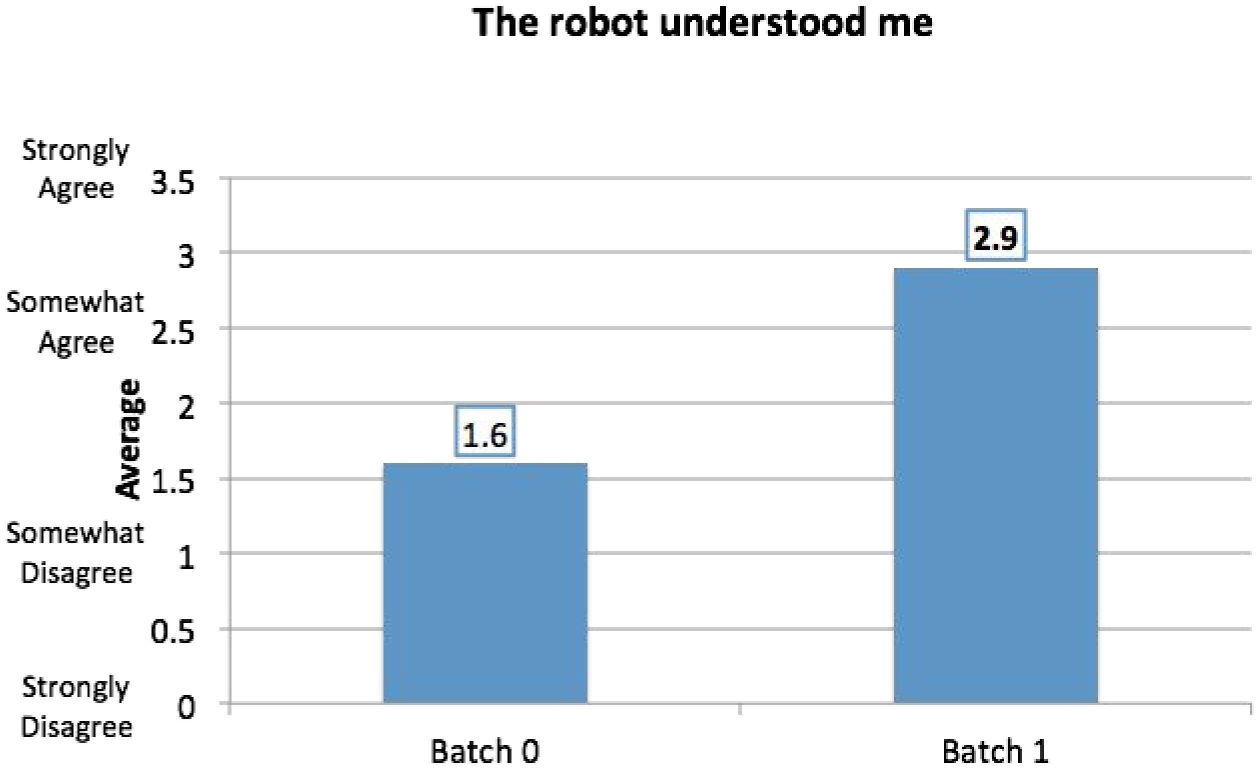
- 10 users with baseline system (no additional training)
- Robot roamed the office for four days
  - 34 conversations with users in the office ended with training goals
  - System re-trained after four days
- 10 users with re-trained system

Segbot Dialog Success

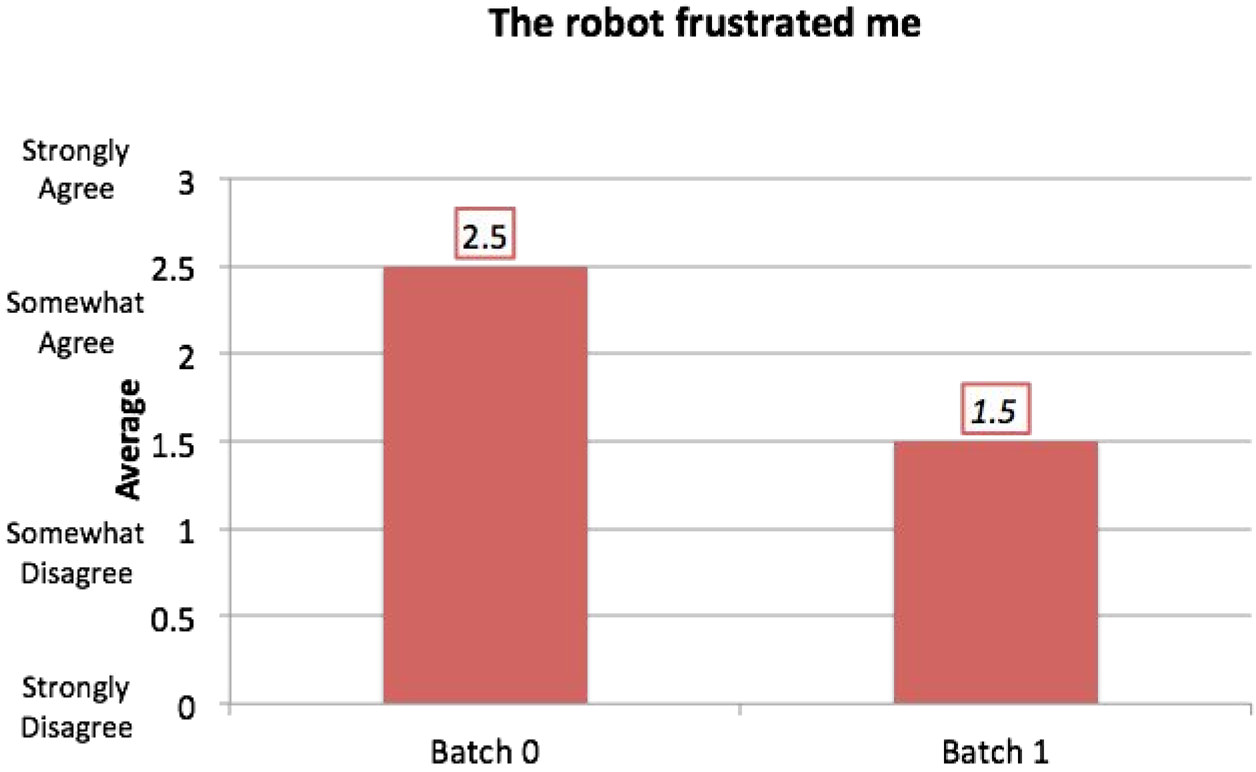




Segbot Survey Responses



Segbot Survey Responses



Findings

- Lexical acquisition reduces dialog lengths for multi-argument predicates like delivery
- Retraining causes users to perceive the system as more understanding
- Retraining leads to less user frustration
- Inducing training data from dialogs allows good language understanding without large, annotated corpora to bootstrap system
- If domain changes or new users with new lexical choices arrive, can adapt on-the-fly

Findings

- Inducing training data from dialogs allows good language understanding without large, annotated corpora to bootstrap system
- If domain changes or new users with new lexical choices arrive, can adapt on-the-fly

“alert me if her heart rate decreases”
“bring me his chart”
“go and get the family”
“scalpel”

“text me when the speaker arrives”
“grab the heavy, green mug”
“lead him to alice’s office”
“get out of the way”

Natural Language Understanding for Robots

Go to Alice’s office and
get the light mug for
the chair.

- Commands that need to be actualized through robot action ✓
- World knowledge about people and the surrounding office space ✓
- Perception information to identify referent object

Background: Language Grounding

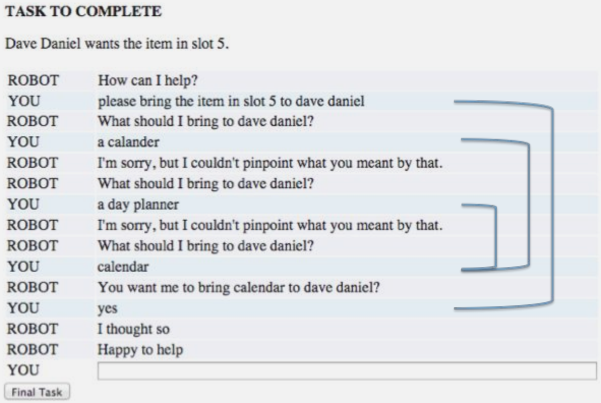
- Some *y* that is light in weight and could be described as a mug
- These predicates are *perceptual* in nature and require using sensors to examine real-world objects for membership

“the light mug”

$$\text{the}(\lambda y.(\text{light}(y) \land \text{mug}(y)))$$

Outline

Learning to Interpret Natural Language Commands through Human-Robot Dialog [Thomason et al. IJCAI 2015]

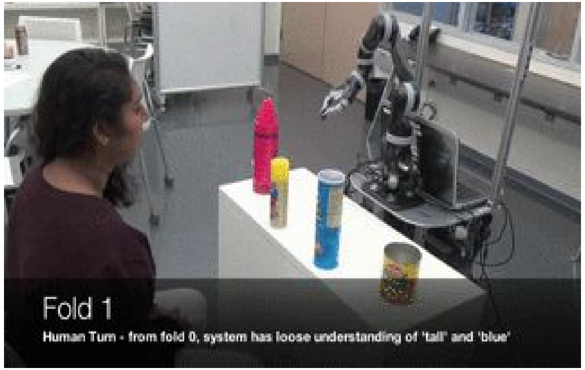


TASK TO COMPLETE

Dave Daniel wants the item in slot 5.

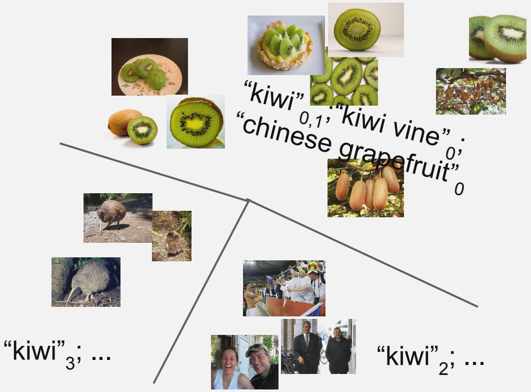
ROBOT	How can I help?
YOU	please bring the item in slot 5 to dave daniel
ROBOT	What should I bring to dave daniel?
YOU	a calander
ROBOT	I'm sorry, but I couldn't pinpoint what you meant by that.
ROBOT	What should I bring to dave daniel?
YOU	a day planner
ROBOT	I'm sorry, but I couldn't pinpoint what 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	

Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy" [Thomason et al. IJCAI 2016]

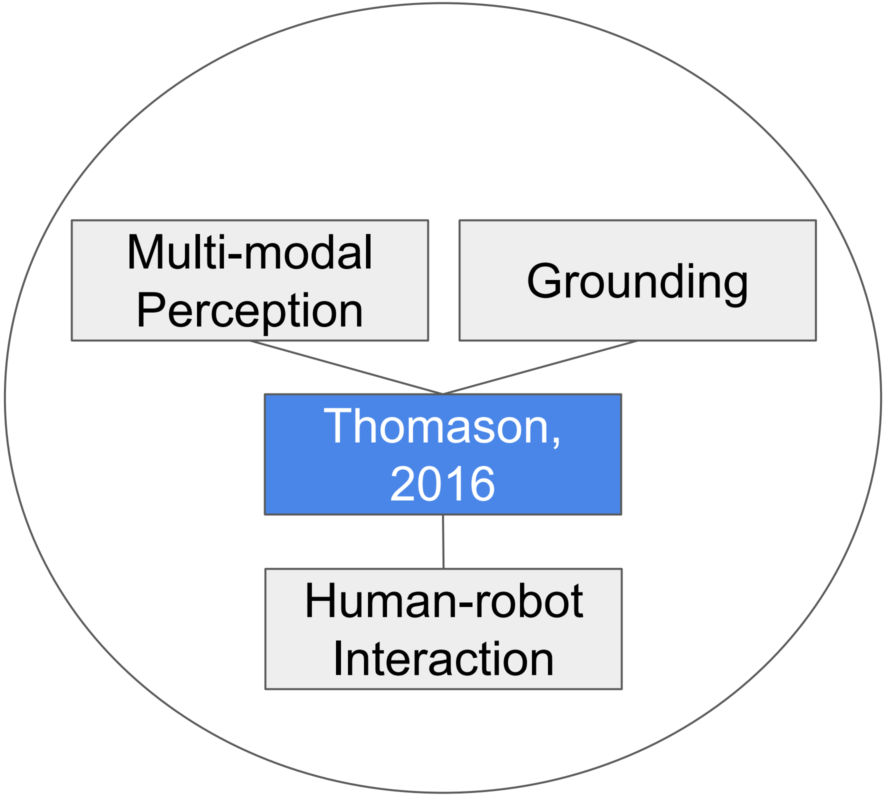


Fold 1
Human Turn - from fold 0, system has loose understanding of 'tall' and 'blue'

Multi-Modal Word Synset Induction [Thomason, Mooney IJCAI 2017]



Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”

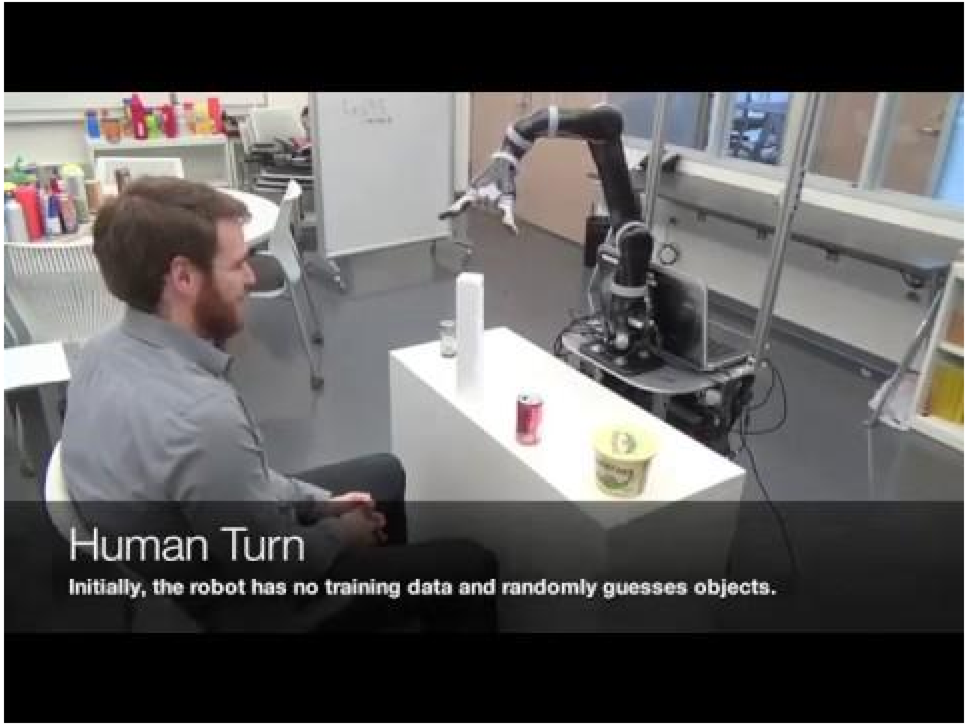


Multi-modal Perception

Grounding

Thomason, 2016

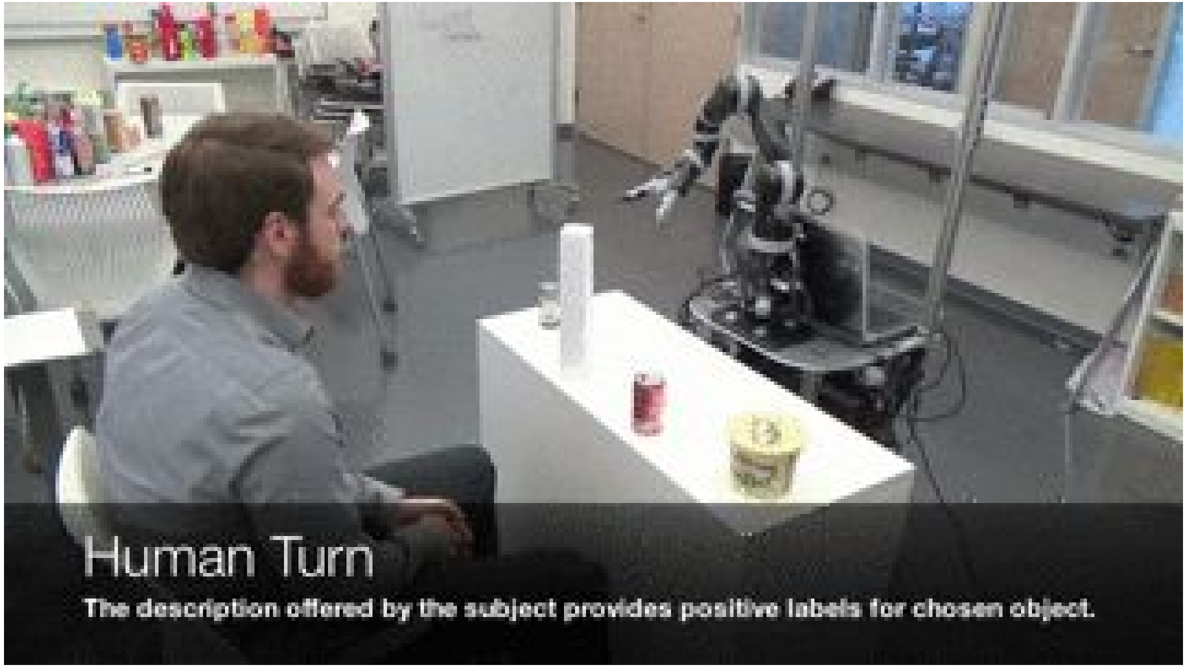
Human-robot Interaction



Human Turn

Initially, the robot has no training data and randomly guesses objects.

“An empty metallic aluminum container”



Human Turn

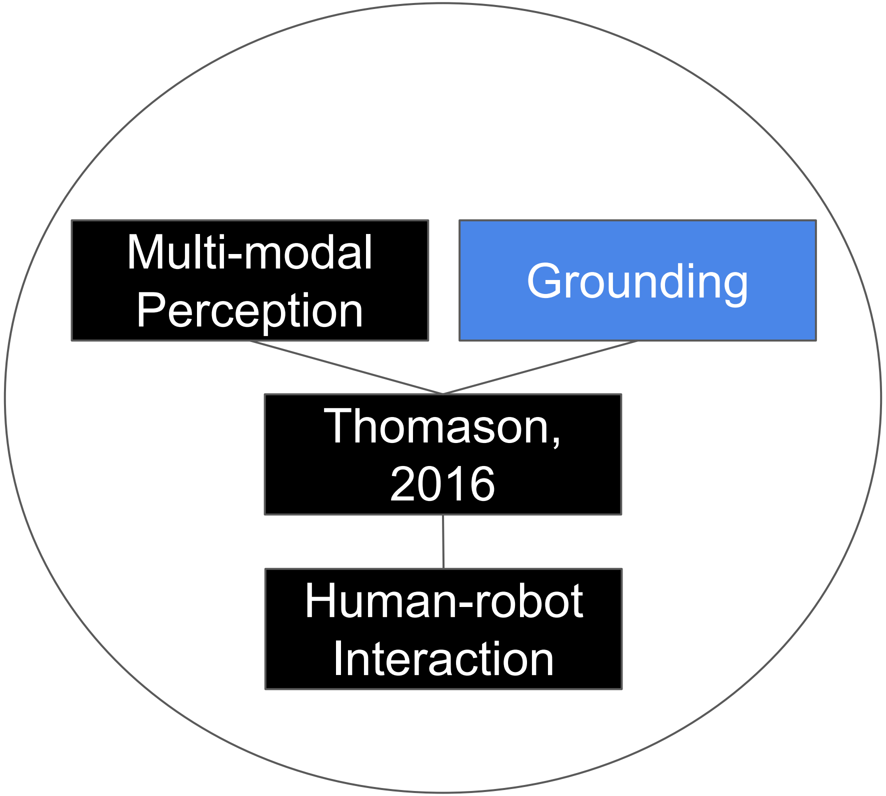
The description offered by the subject provides positive labels for chosen object.

“An empty metallic aluminum container”

Human Turn
Initially, the robot has no training data and randomly guesses objects.

Robot makes guesses until human confirms it found the right object.

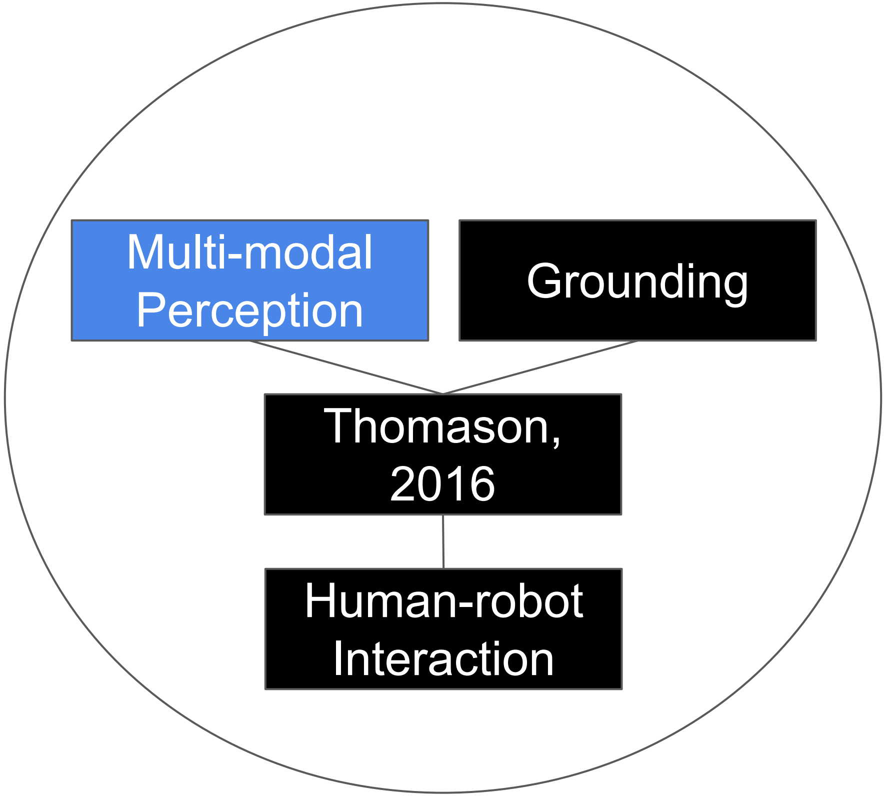
Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”



Grounding

- Mapping from expressions like “light mug” to an object in the real world is the *symbol grounding problem* [Harnad, 1990]
- *Grounded language learning* aims to solve this problem
    - Essential for robots to perform object retrieval tasks (e.g. “bring me his chart”; “grab the heavy, green mug”)
- Loads of work connecting language to machine vision [Roy, 2002; Matuszek, 2012; Krishnamurthy, 2013; Christie, 2016]
- Some work connecting language to other perception, such as audio [Kiela, 2015]
- We ground words in more than just vision

Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”



Multi-modal Perception

Grounding

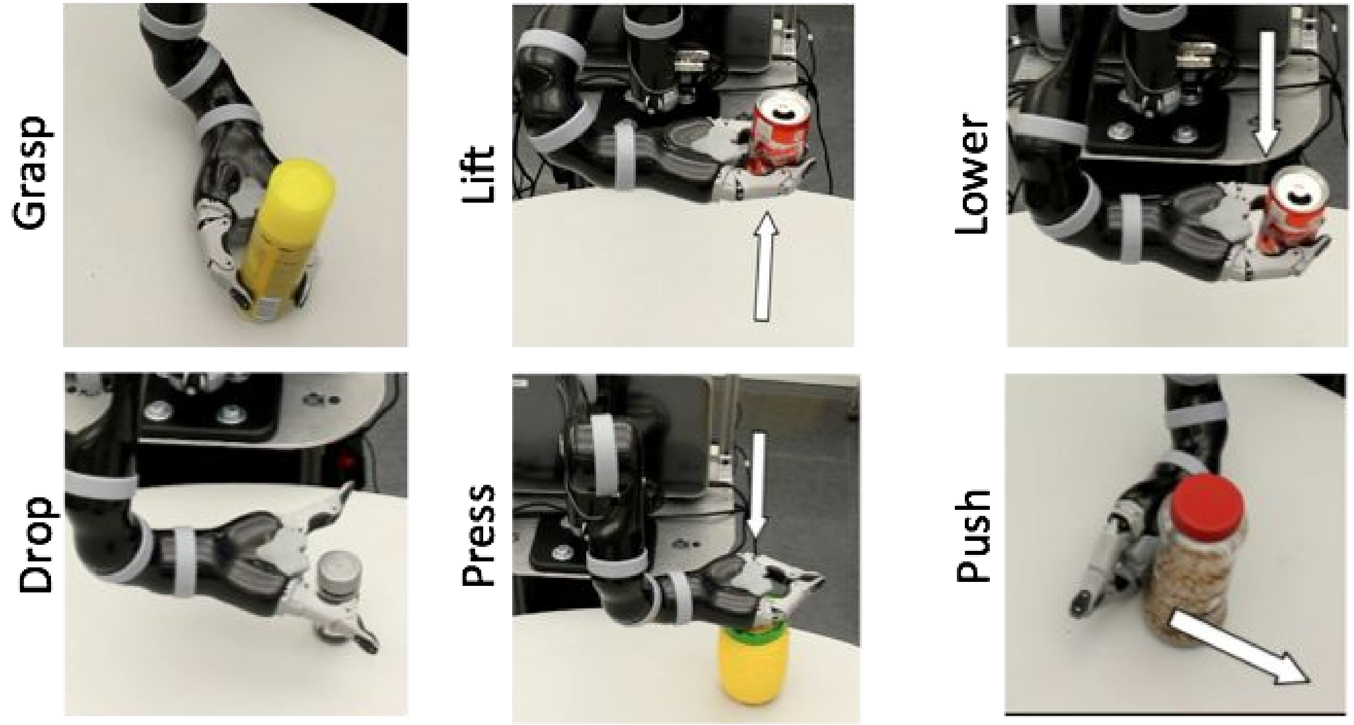
Thomason, 2016

Human-robot Interaction

Multi-Modal Perception

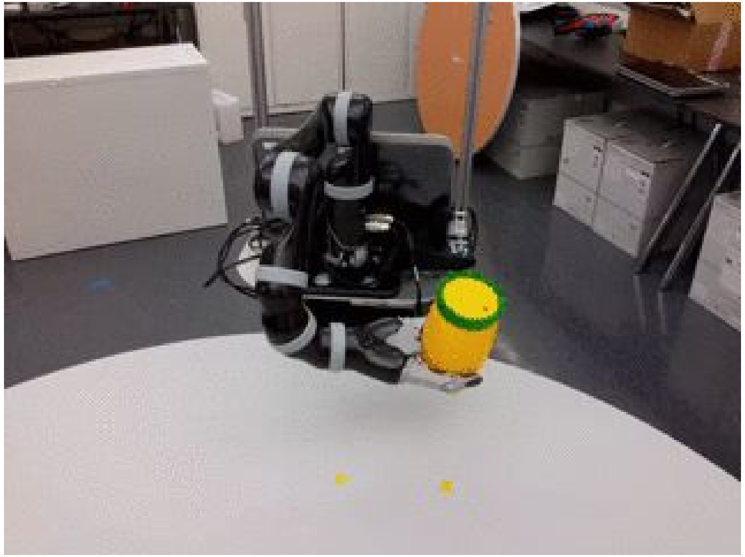
- For every object, perform a set of exploratory behaviors (with robotic arm) [Sinapov, 2016]
- Gather audio signal from microphone and, proprioceptive and haptic information from arm motors
- “Look” is just one way to explore; gathers deep features, color histograms, and fast point feature histograms
- Feature representation of each object has many sensorimotor *contexts*
- Context is a combination of an exploratory behavior and associated sensory modality

Multi-Modal Perception



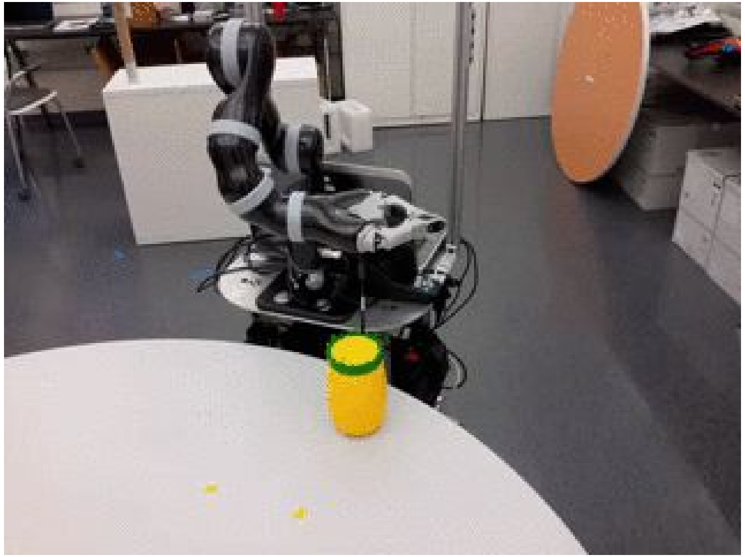
Multi-Modal Perception

lift, hold, lower



drop

Multi-Modal Perception



press

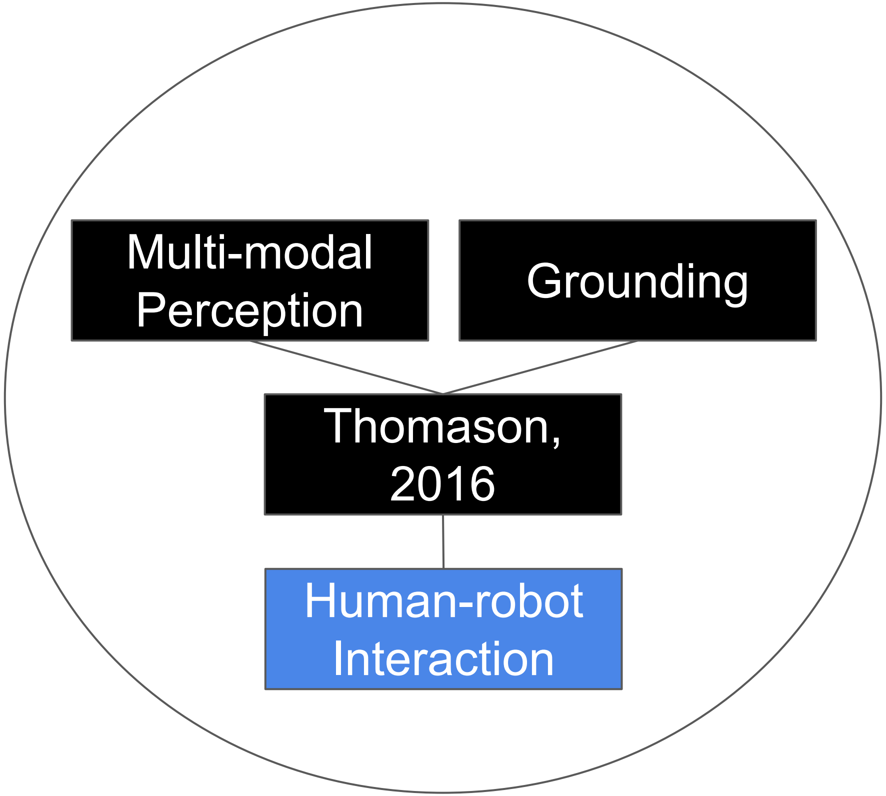
push

Multi-Modal Perception

Multi-Modal Perception

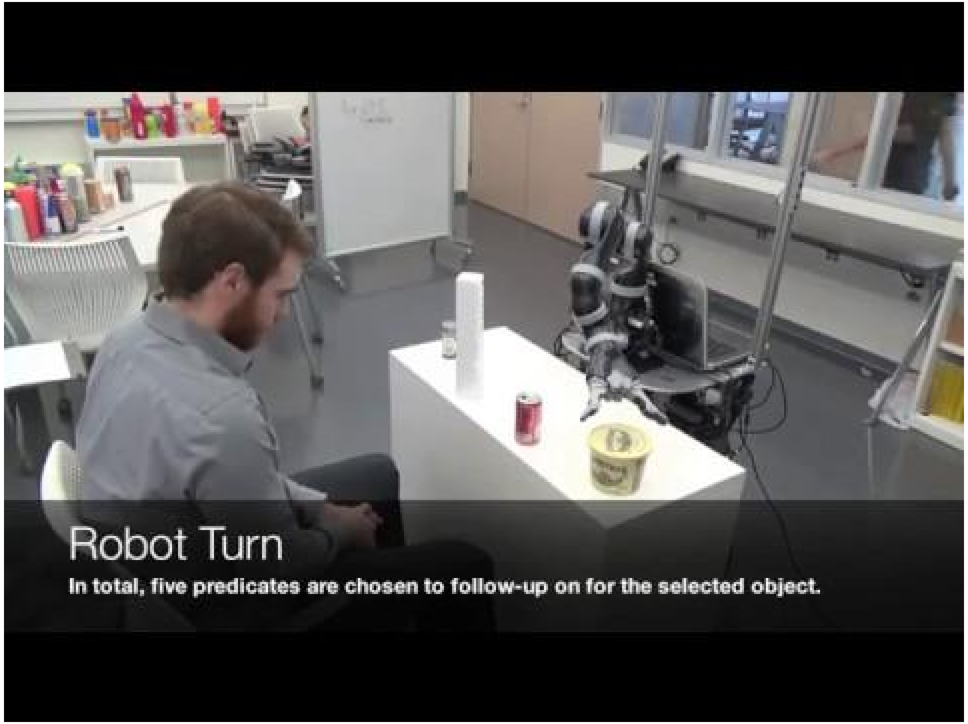
- Still need language labels for objects
- Annotating each object with every possible descriptor is unrealistic and boring
- Can’t use online annotators to get non-visual descriptors like “heavy”, “full”, or “rattles”; objects need to be interacted with in person
- Instead, we introduce a human-in-the-loop for learning
- In a game!

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



Human-robot Interaction

- Past work has used “I, Spy”-like games to gather grounding annotations from users [Parde 2015]
  - Humans like playing with robots (for a while), especially if the robots get smarter
- Human offers natural language description of object
- Robot strips stopwords and treats remaining words as predicate labels
- On robot’s turn, use predicates to determine best way to describe target object
- Ask for explicit yes/no on whether some predicates apply to target (e.g. “would you use the word ‘heavy’ to describe this object?”)



Robot Turn

In total, five predicates are chosen to follow-up on for the selected object.

“Would you use the word ‘half-full’ when describing this object?”

“Yes”

Robot Turn
A follow-up dialog gives additional positive/negative labels for predicates.

R: “Would you use the word ‘half-full’ when describing this object?”

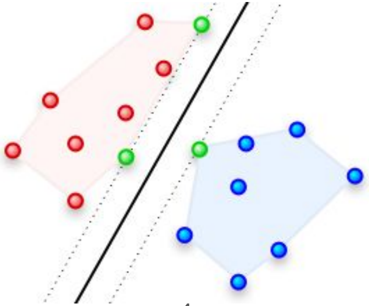
H: “Yes”

Building Perceptual Classifiers

- Get positive labels from human descriptions of target objects
- Get positive and negative labels from yes/no answers to specific predicate questions
- Build SVM classifiers for each sensorimotor context given positive and negative objects for each predicate
- Predicate classifier is linear combination of context SVMs
- Weight each SVM’s decision by kappa agreement with users using leave-one-out x-val over objects

Building Perceptual Classifiers

Sensorimotor context SVM



"empty"?

Behavior / Modality	color	...	audio	haptics
look	0.02		-	-
...	...	...	...	...
lift	-	...	-0.04	0.8
drop	-	...	0.4	0.02

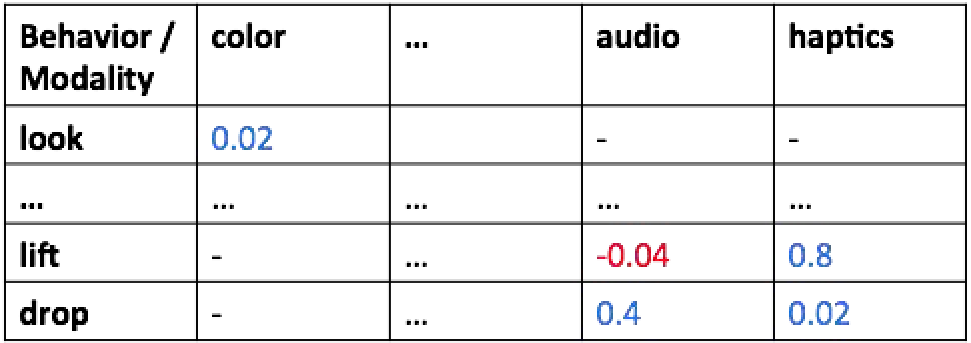
Prediction gives sign in {-1, 1}

Agreement with human labels under leave-one-out xval gives magnitude

Building Perceptual Classifiers

$$(0.02 + \ldots + (-0.04) + 0.8 + 0.4 + 0.02) / 18 = 0.076$$

“empty”?



Behavior / Modality	color	...	audio	haptics
look	0.02		-	-
...	...	...	...	...
lift	-	...	-0.04	0.8
drop	-	...	0.4	0.02

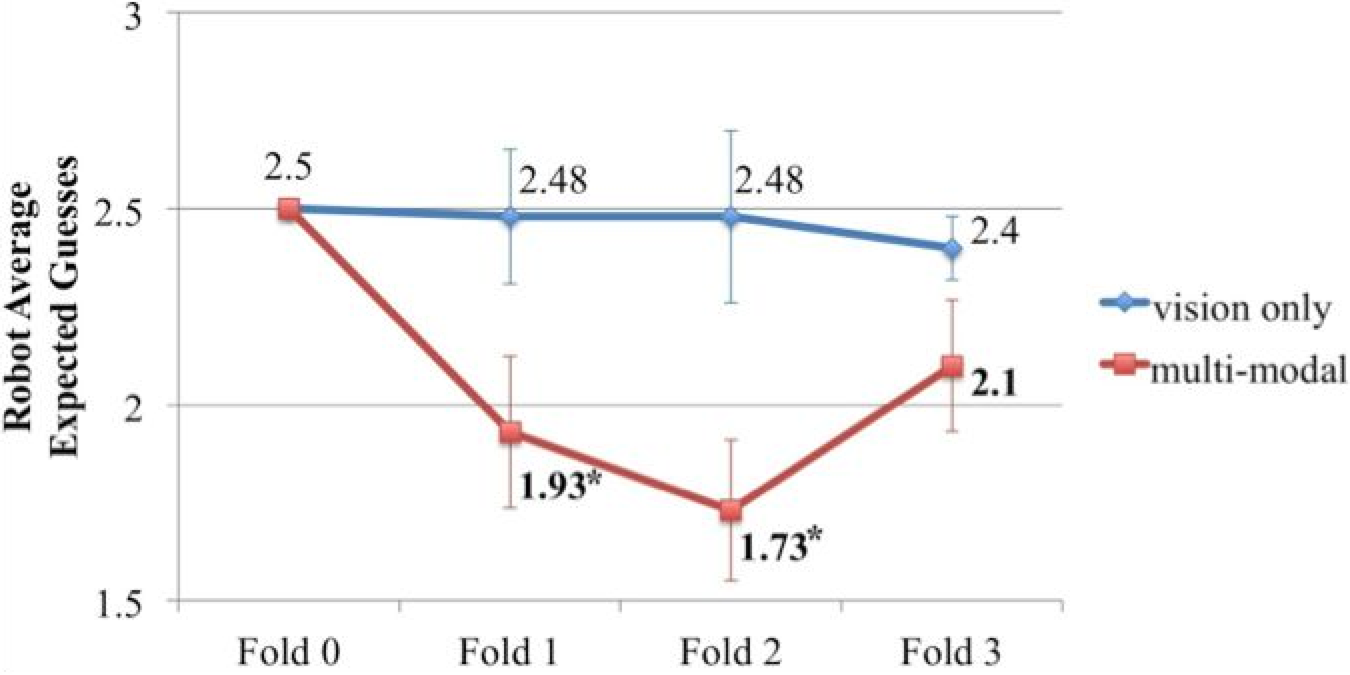
Prediction gives sign in {-1, 1}

Agreement with human labels under leave-one-out xval gives magnitude

Experiments

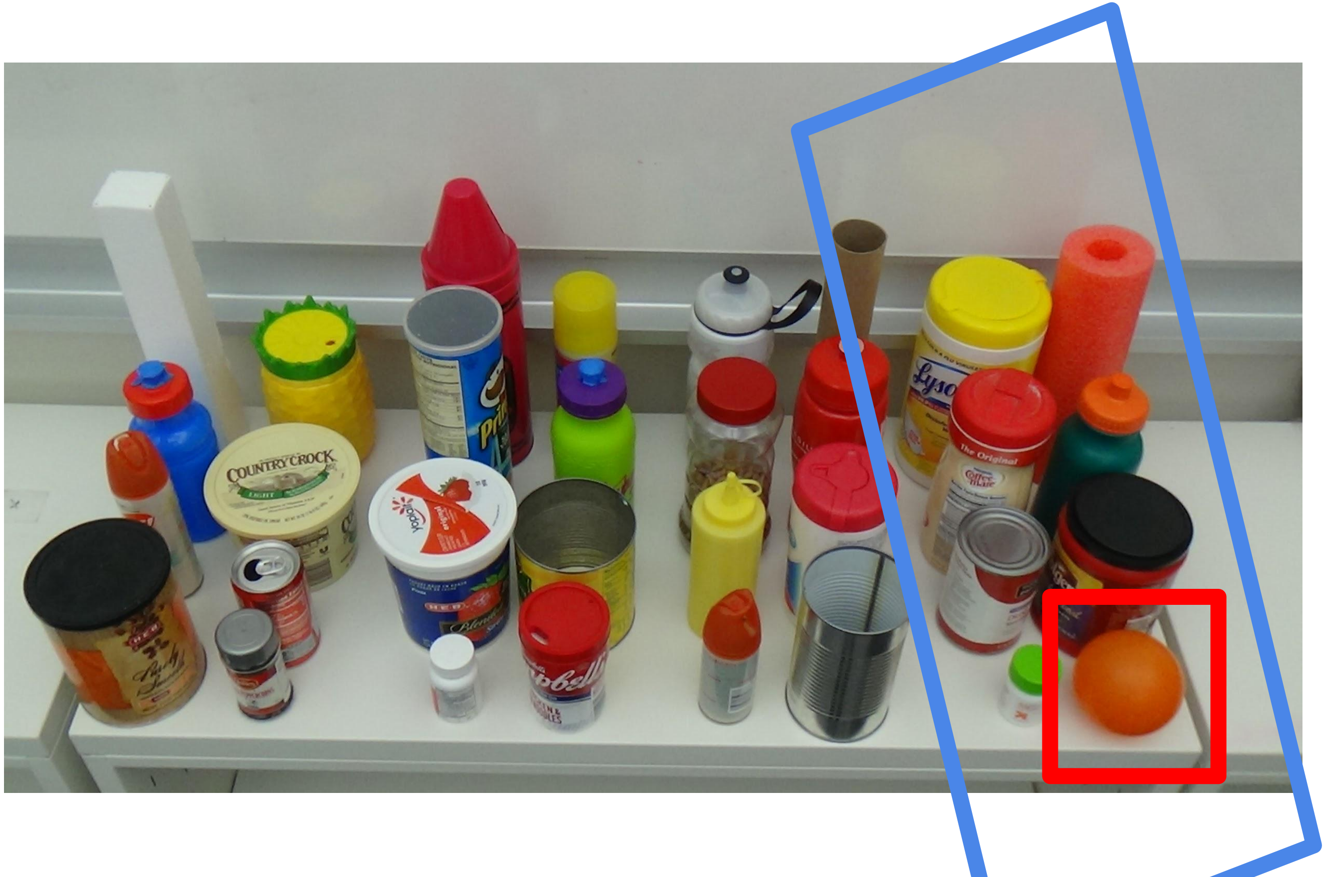
- 32 objects split into 4 folds of 8 objects each
- Games played with 4 objects at a time
- Two systems: **vision only** and **multi-modal**; former only uses *look* behavior
- Each participant played 4 games, 2 with each system (single blind), such that each system saw all 8 objects of the fold
- After each fold, systems' predicate classifiers retrained given new labels
- Measure game performance; classifiers always seeing novel objects during evaluations

Results for Robot Guesses



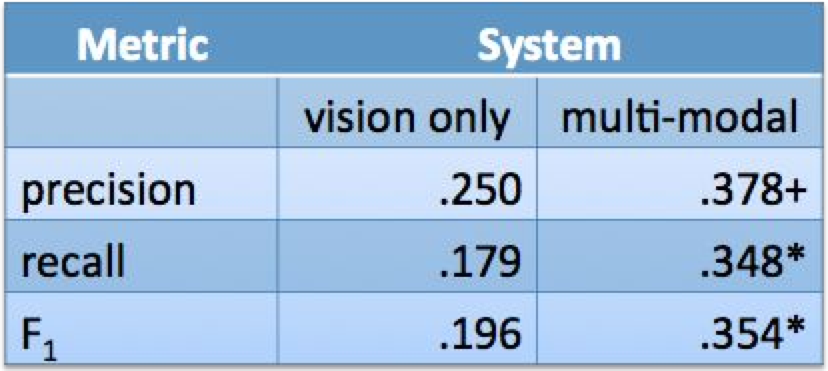
Bold: Lower than fold 0 average. *: Lower than vision only baseline

Results for Robot Guesses



Results for Predicate Agreement

- Leave-one-object-out cross validation across predicate labels on objects (74 total learned)

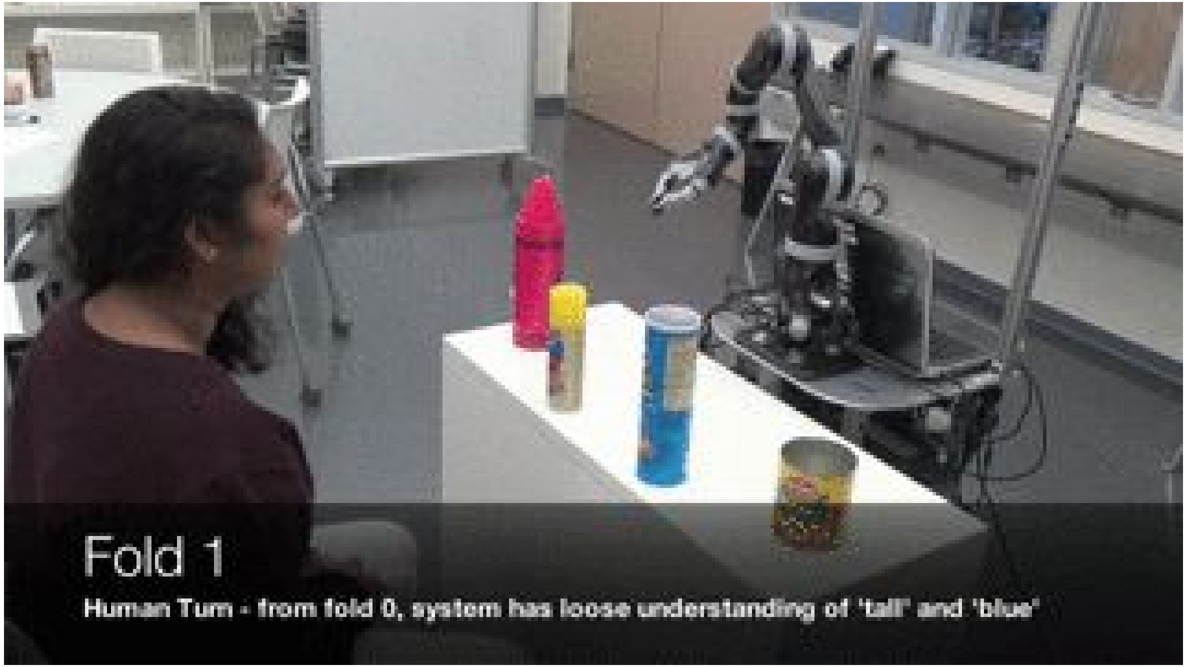


Metric	System	
	vision only	multi-modal
precision	.250	.378+
recall	.179	.348*
F_1	.196	.354*

- *: significantly greater with $p < 0.05$
- +: trending greater with $p < 0.1$

Correlations to Physical Properties

- Calculated Pearson’s *r* between predicate decisions in [-1, 1] and object height/weight
- **vision only** system learns no predicates with $p < 0.05$ and $|r| > 0.5$
- **multi-modal** system learns several correlated predicates:
    - “tall” with height (r = 0.521)
    - “small” against weight (r = -0.665)
    - “water” with weight (r = 0.549)



Fold 1

Human Turn - from fold 0, system has loose understanding of 'tall' and 'blue'

“A tall blue cylindrical container”

Findings

- Auditory, haptic, and proprioceptive senses help understand words humans use to describe objects
- Some predicates assisted by multi-modal
  - “tall”, “wide”, “small”
- Some can be impossible without multi-modal
  - “half-full”, “rattles”, “empty”

Natural Language Understanding for Robots

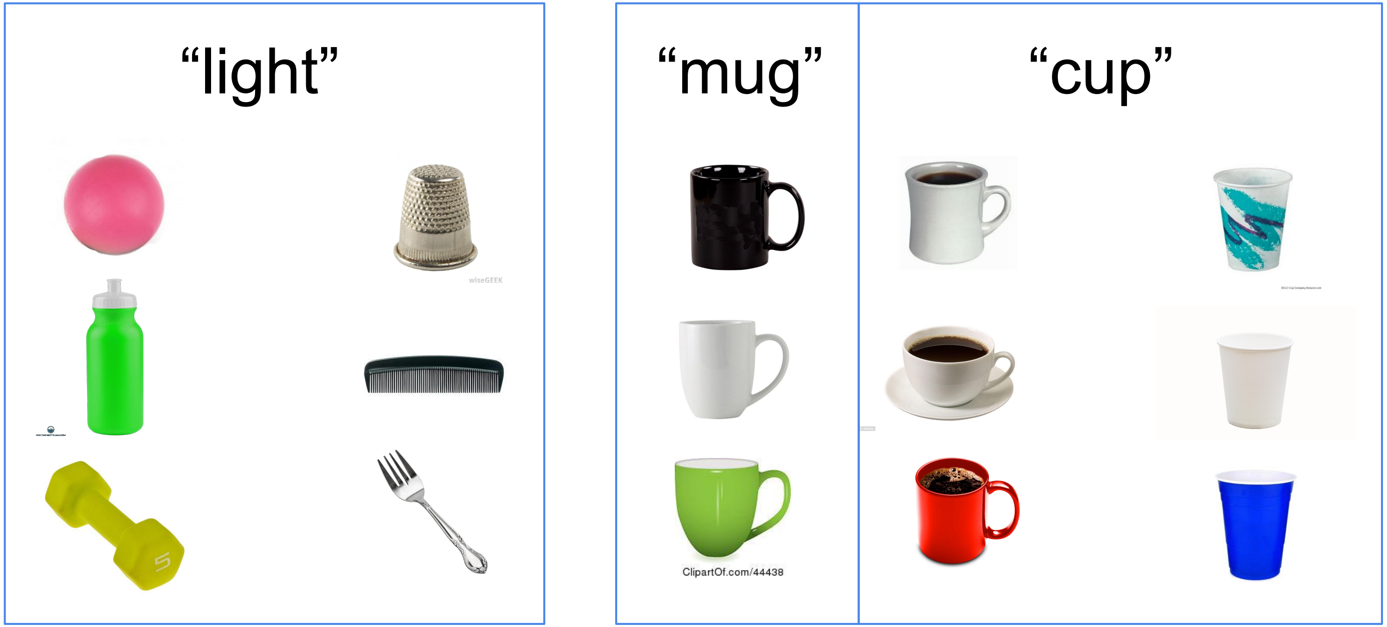
> Go to Alice’s office and get the light mug for the chair.

- Commands that need to be actualized through robot action
- World knowledge about people and the surrounding office space
- Perception information to identify referent object ✓
    - But we don’t handle different senses of light

Background: Language Grounding

word

instances

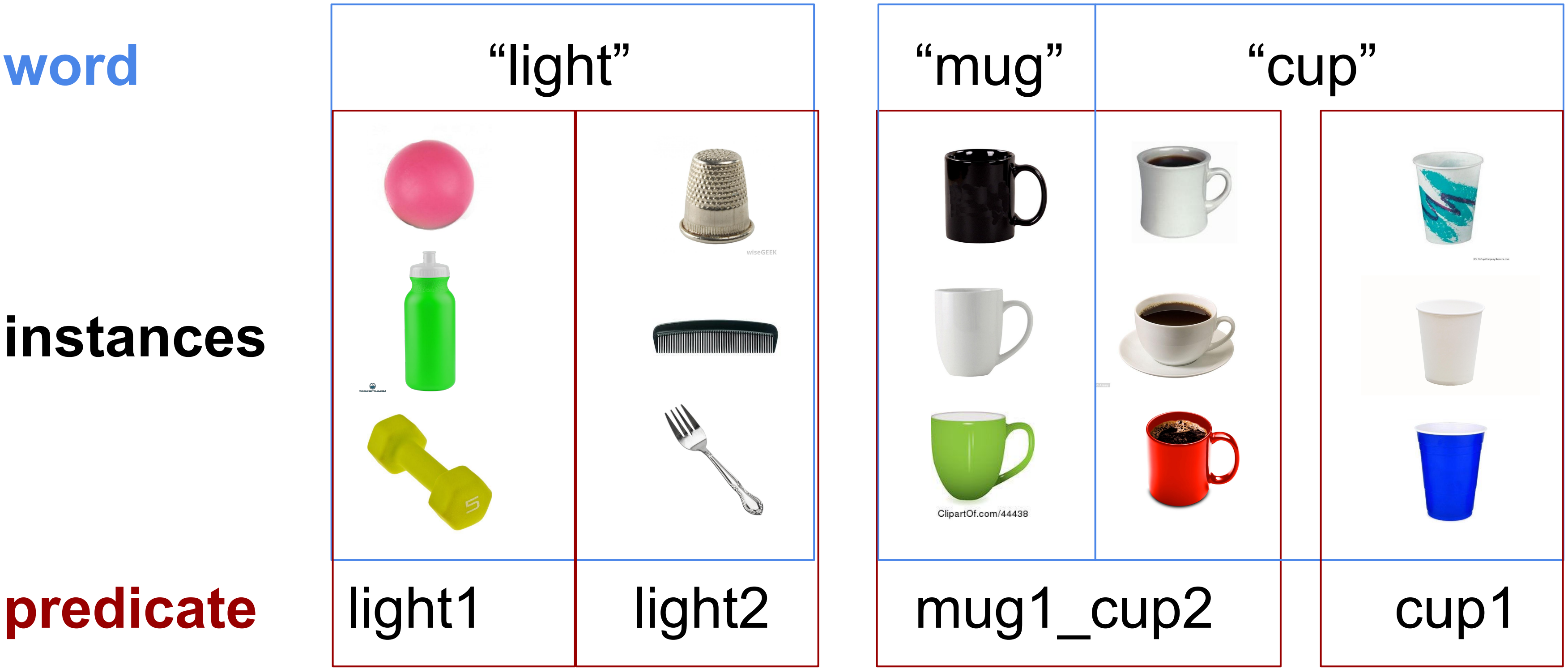


“light”

“mug”

“cup”

Background: Language Grounding



word

“light”

“mug”

“cup”

instances

predicate

light1

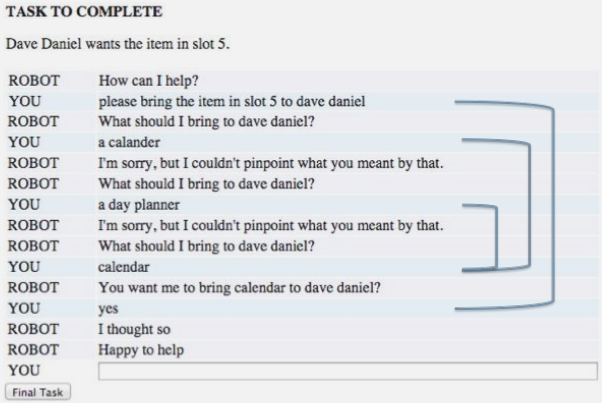
light2

mug1_cup2

cup1

Outline

Learning to Interpret Natural Language Commands through Human-Robot Dialog [Thomason et al. IJCAI 2015]

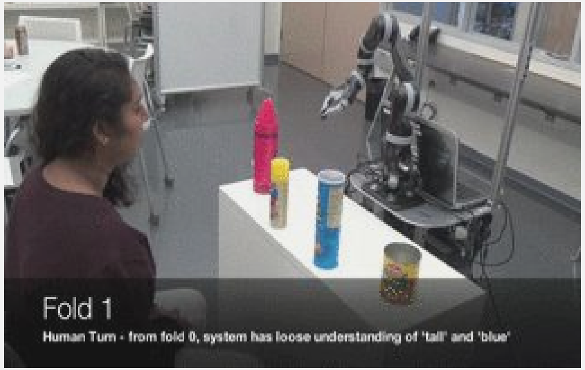


TASK TO COMPLETE

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	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	

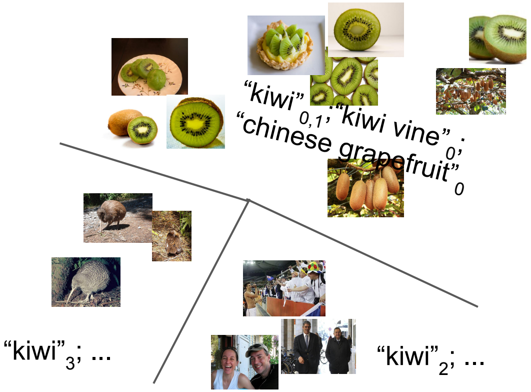
Learning Multi-Modal Grounded Linguistic Semantics by Playing "I Spy" [Thomason et al. IJCAI 2016]



Fold 1

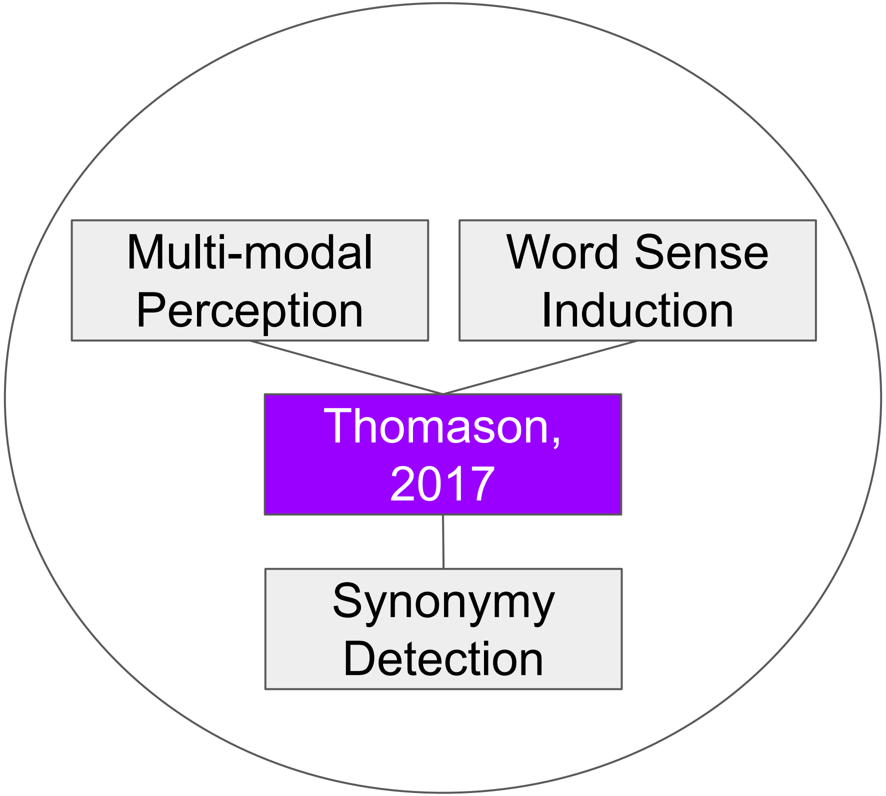
Human Turn - from fold 0, system has loose understanding of 'tall' and 'blue'

Multi-Modal Word Synset Induction [Thomason, Mooney IJCAI 2017]

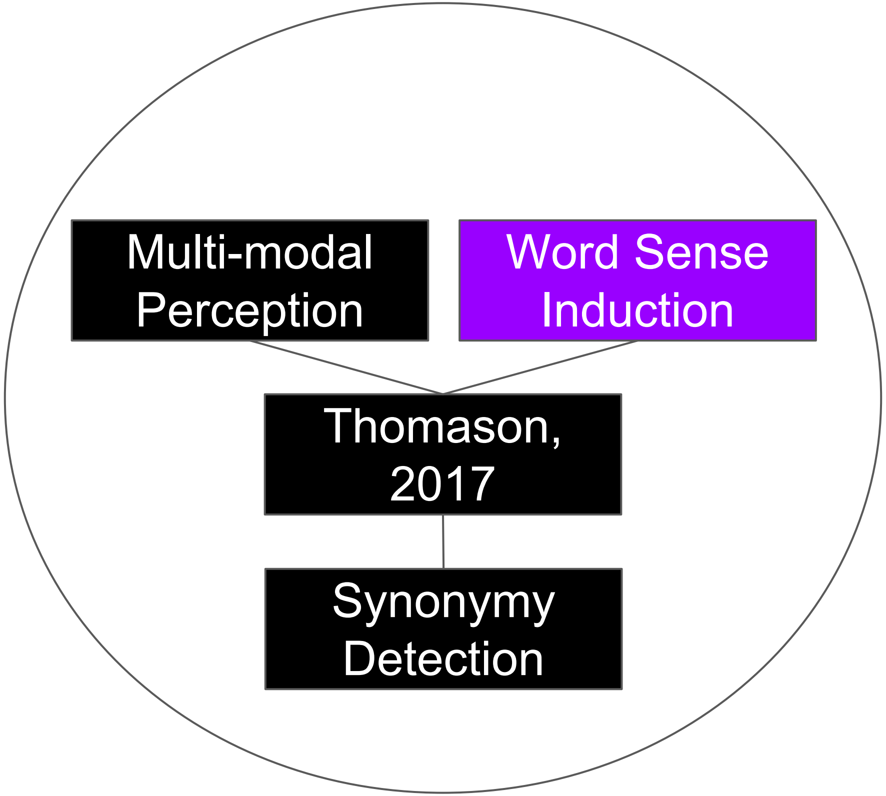


“kiwi”$_{0,1}$; “kiwi vine”$_{0}$; “chinese grapefruit”$_{0}$
“kiwi”$_{3}$; ...
“kiwi”$_{2}$; ...

Multi-Modal Word Synset Induction



Multi-Modal Word Synset Induction



Multi-modal Perception

Word Sense Induction

Thomason, 2017

Synonymy Detection

Word Sense Induction

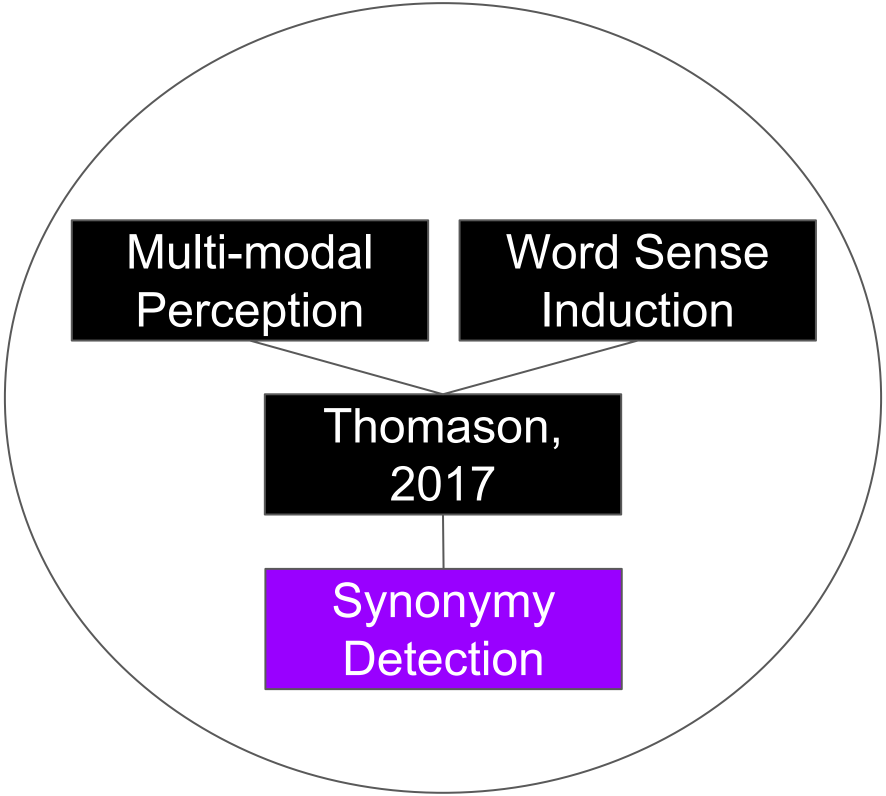
- Task of discovering word senses [Pedersen and Bruce, 1997]
- “Bat”
  - Baseball, animal

- “Light”
  - Weight, color

- “Kiwi”
  - Fruit, bird, people

- Represent instances as vectors of their context; cluster to find senses
  - [Yarowsky, 1995; Pedersen and Bruce, 1997; Schutze, 1998; Bordag, 2006; Navigli, 2009; Manandhar et al., 2010; Di Marco and Navigli, 2013]

Multi-Modal Word Synset Induction



Multi-modal Perception

Word Sense Induction

Thomason, 2017

Synonymy Detection

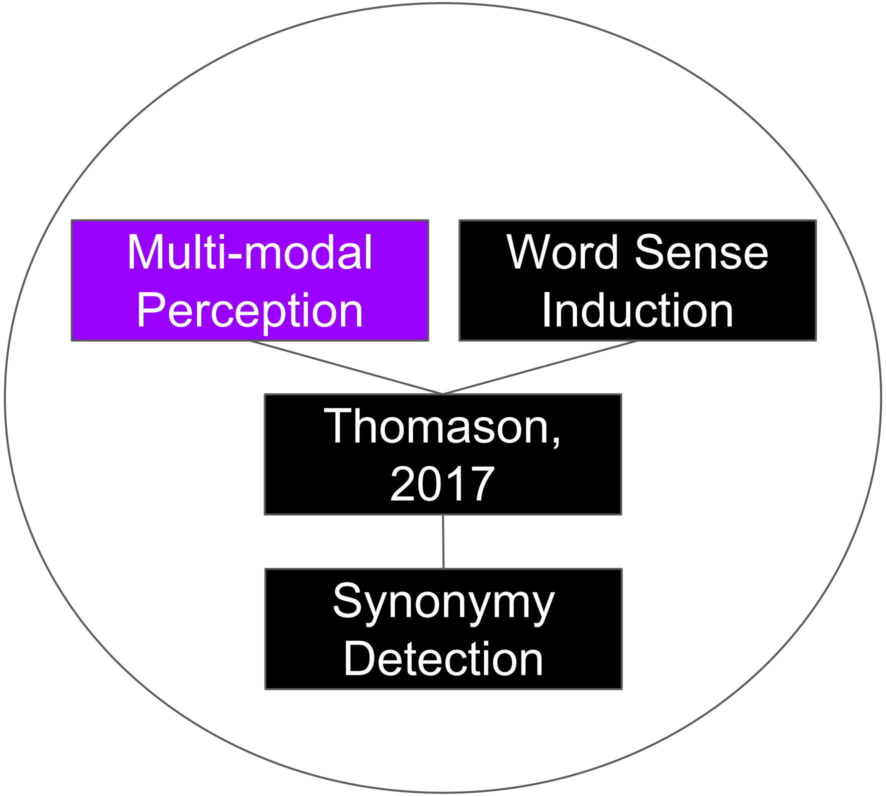
Synonymy Detection

- Given words or word senses, find synonyms
- “Ball” and “sphere”
- “Round” and “circular”
- “Kiwi” and “New Zealander” (for one sense of “kiwi”)
- Represent instances as vectors of their context; cluster means to find synonyms
    - Related to synonym detection [Turney, 2001] and lexical substitution [McCarthy and Navigli, 2009]

Word Sense Induction + Synonymy Detection

- First finding senses, then merging those senses through synonymy detection
- We call this *synset induction*, the task of finding synonymous sets of word senses
- Synsets used in WordNet [Fellbaum, 1998] and analogous ImageNet [Deng et al., 2009] corpora
    - Represent hierarchical collections of synonymous noun phrases
    - e.g. “kiwi”, “chinese grapefruit”, “kiwi vine”

Multi-Modal Word Synset Induction



Multi-modal Perception

- Can use more than text to contextualize a word
- Pictures depicting the word or phrase give visual information

“about 70% of bat species are insectivores”

“most of the oldest known, definitely identified bat fossils were already very similar to modern microbats”

“a baseball bat is divided into several regions”

“hickory has fallen into disfavor over its greater weight, which slows down bat speed”

Dataset

- Gather many leaf-level synsets (6710) and images from ImageNet
- Get a mix of noun phrase types (8426 total)
  - Many past works assume all words are polysemous (e.g. [Loeff et al., 2006; Saenko and Darrell, 2008])

Noun phrase relationships			
synonymous	**polysemous**	**both**	**neither**
4019	804	1017	2586

- Provides “gold” synsets we aim to construct from image-level instances

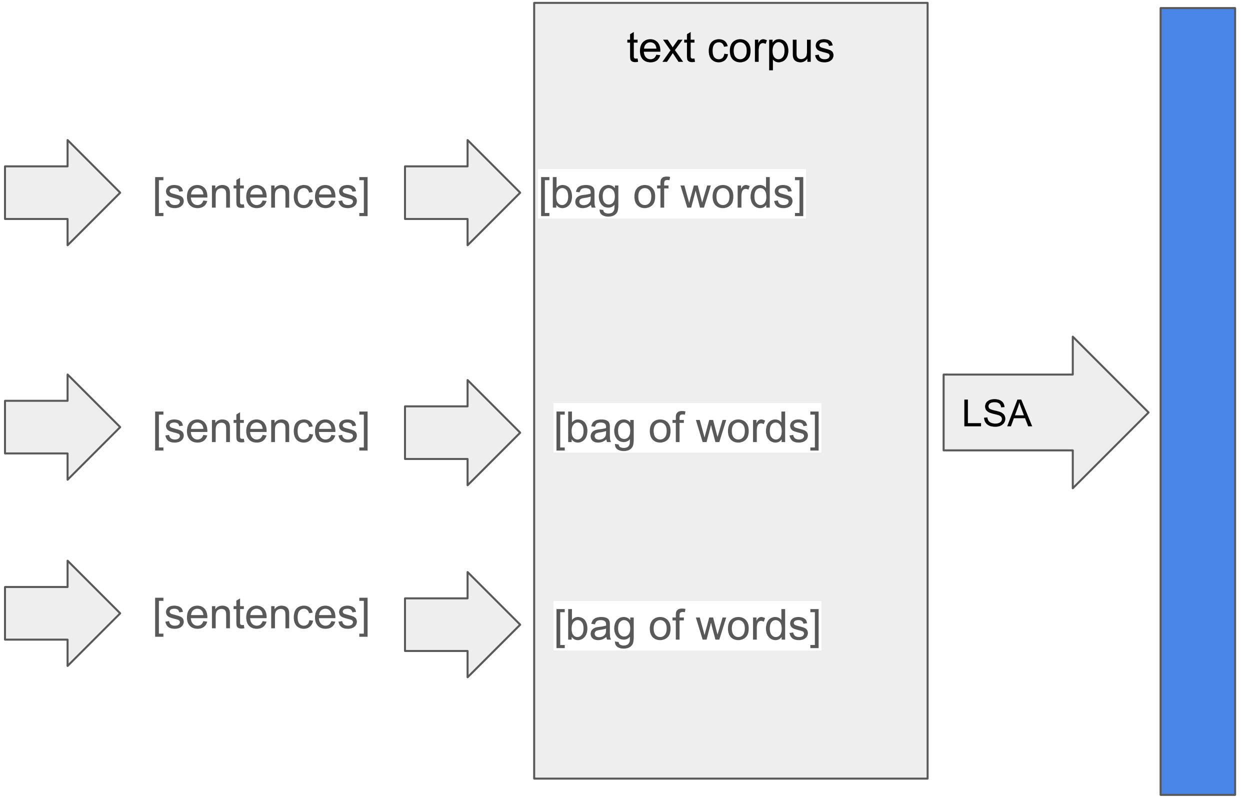
Dataset

- Use reverse-image search to find webpages of text for each image
    - Get textual features and perform clustering in multi-modal space

Reverse image search
Get sentences of webpage

“about 70% of bat species are insectivores”
“most of the oldest known, definitely identified bat fossils were already very similar to modern microbats”
….

Dataset



[sentences]

[bag of words]

[sentences]

[bag of words]

[sentences]

[bag of words]

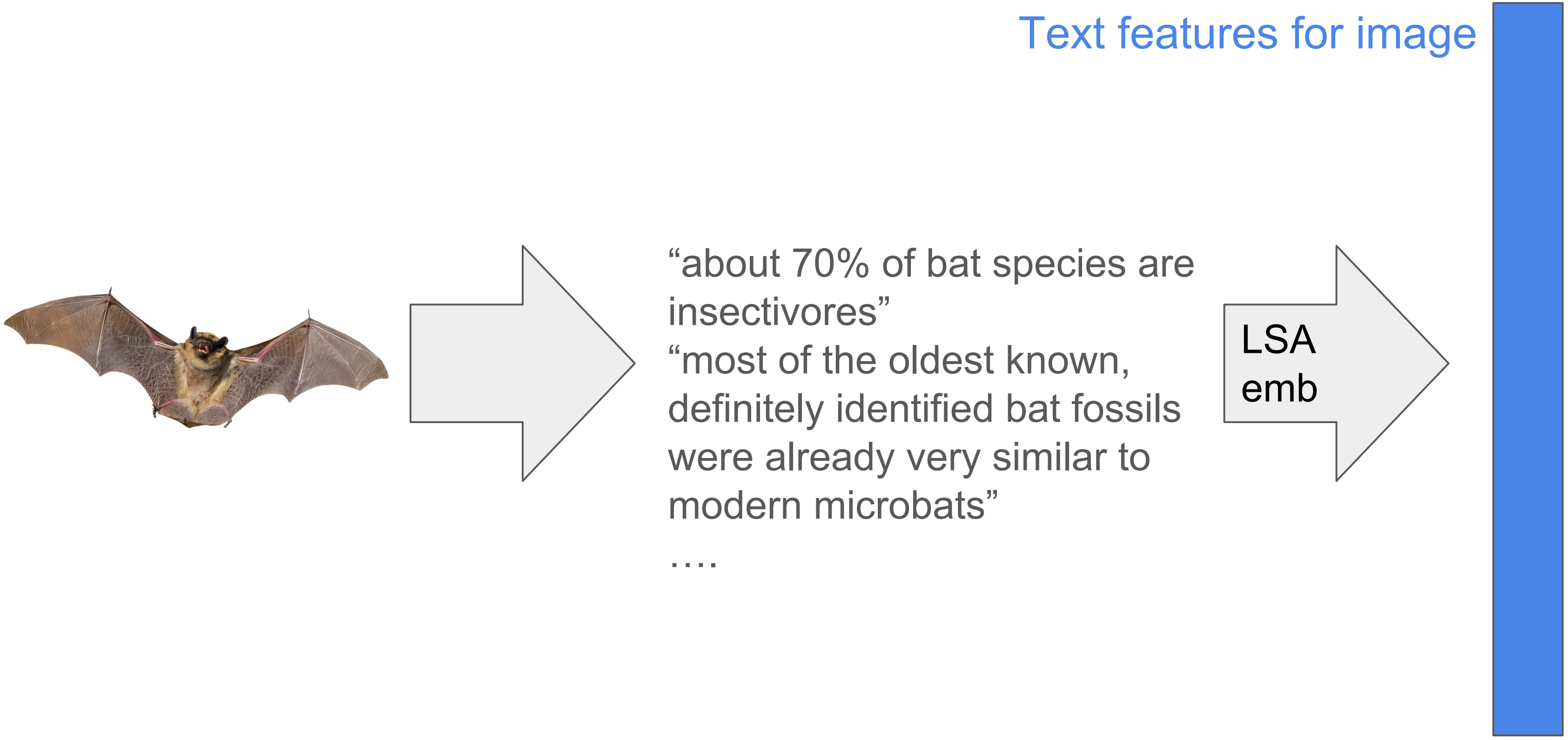
text corpus

LSA

256-dimensional

text feature space

Dataset

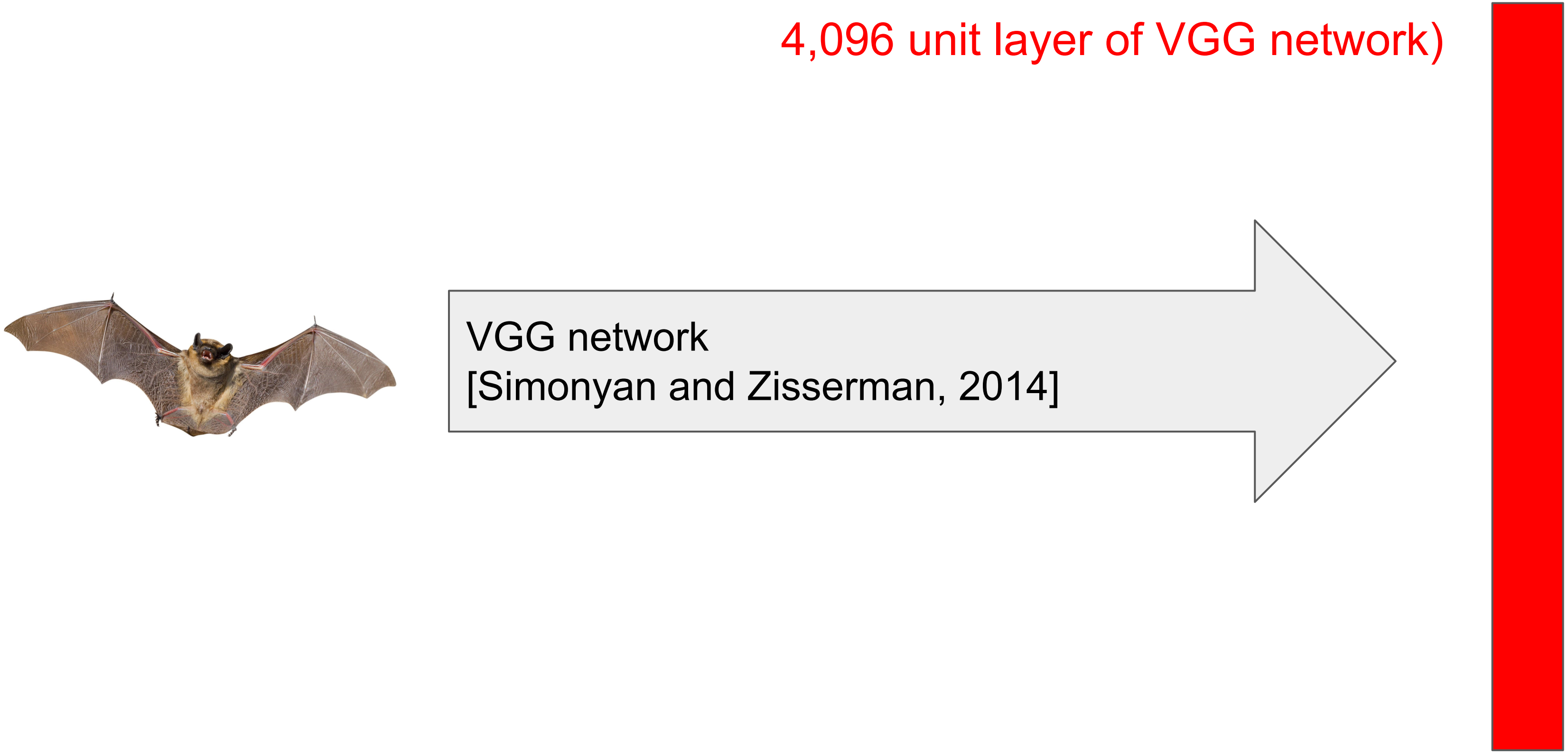


Text features for image

“about 70% of bat species are insectivores”
“most of the oldest known, definitely identified bat fossils were already very similar to modern microbats”
….

LSA emb

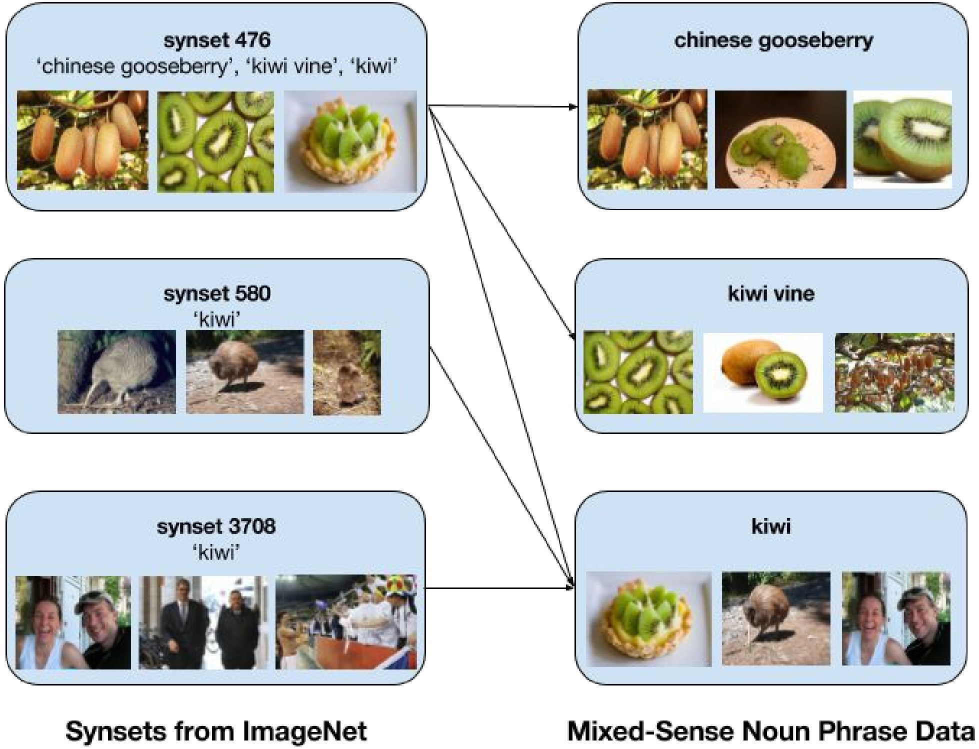
Dataset



Visual features for image (penultimate 4,096 unit layer of VGG network)

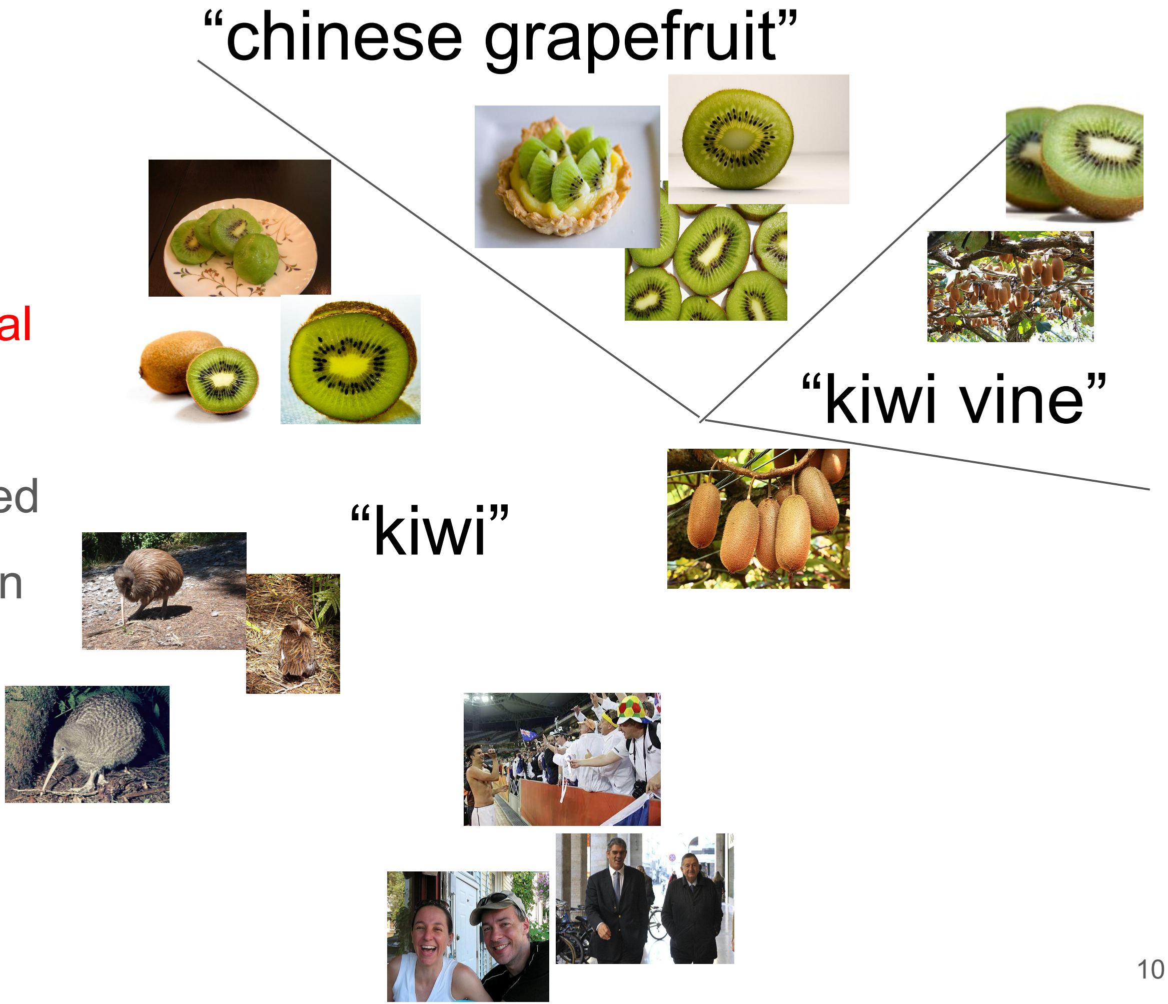
VGG network
[Simonyan and Zisserman, 2014]

Dataset


synset 476
'chinese gooseberry', 'kiwi vine', 'kiwi'
synset 580
'kiwi'
synset 3708
'kiwi'
chinese gooseberry
kiwi vine
kiwi
Synsets from ImageNet
Mixed-Sense Noun Phrase Data

Dataset

- Each image has associated text and visual features
- Feature embeddings used to find distances between image observations

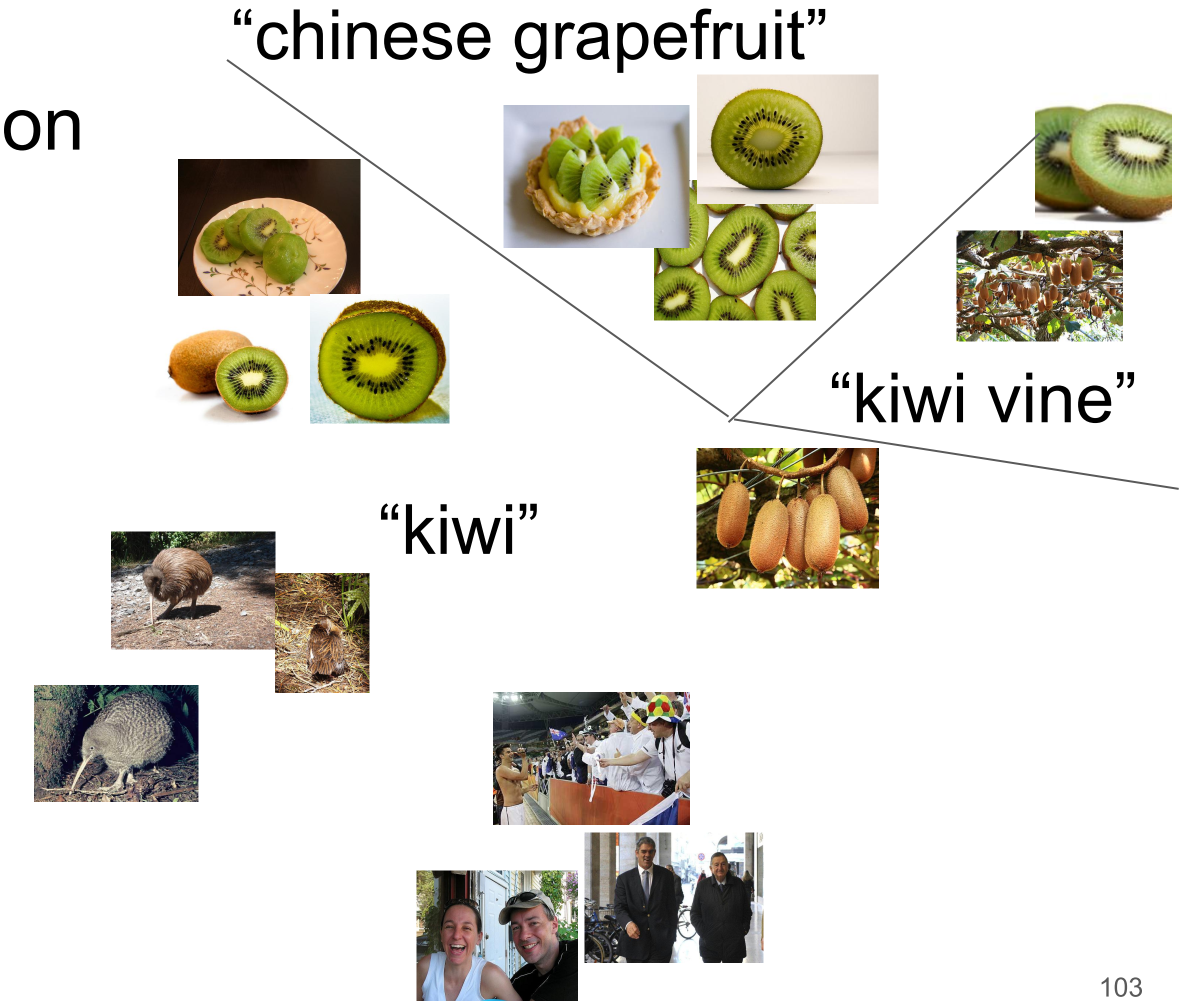


Goal

- Construct ImageNet-like synsets from images labeled with just noun phrase
- First perform word-sense induction on mixed-sense noun phrase inputs
- Given induced word senses, perform synonymy detection to form synsets
- Compare constructions considering text-only, visual-only, and multi-modal spaces
- For multi-modal space, interpolate distance calculations in text and visual spaces

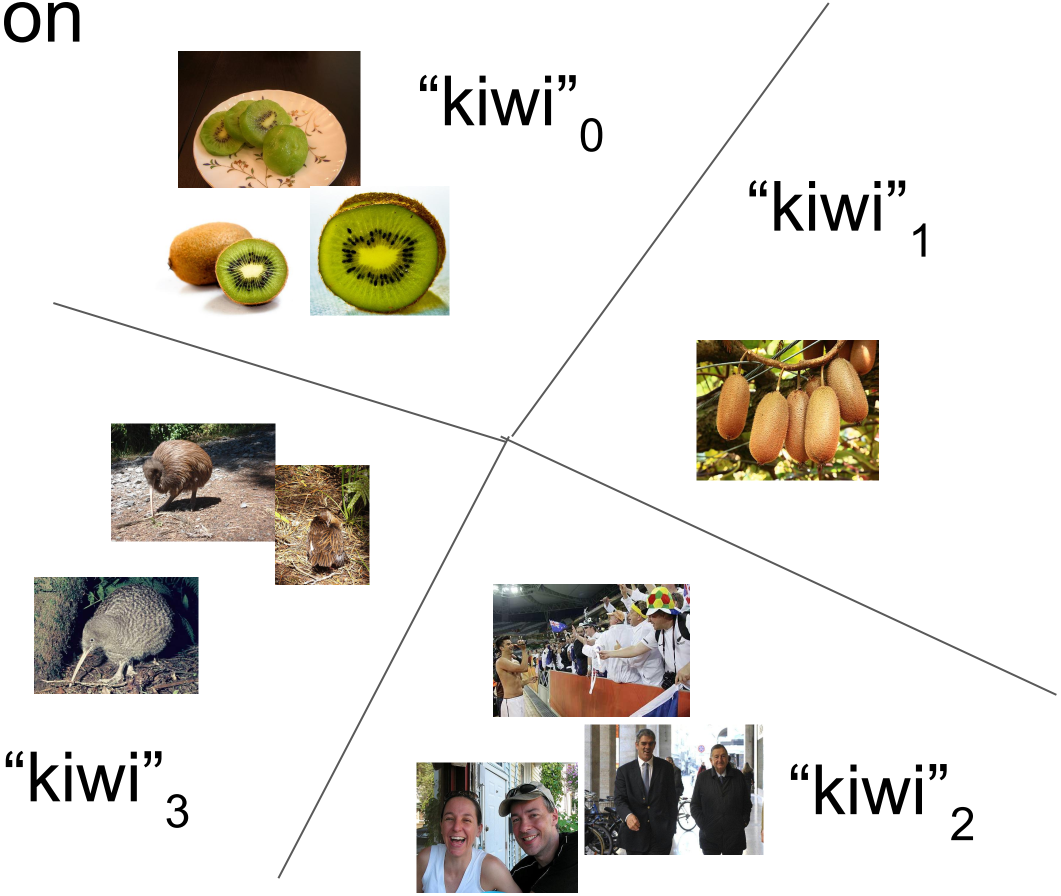
Word Sense Induction

- For every noun phrase, we perform k-means clustering to find senses
- Determine k by the gap statistic [Tibshirani et al., 2001]



Word Sense Induction

- For every noun phrase, we perform k-means clustering to find senses
- Determine k by the gap statistic [Tibshirani et al., 2001]



“kiwi”$_0$

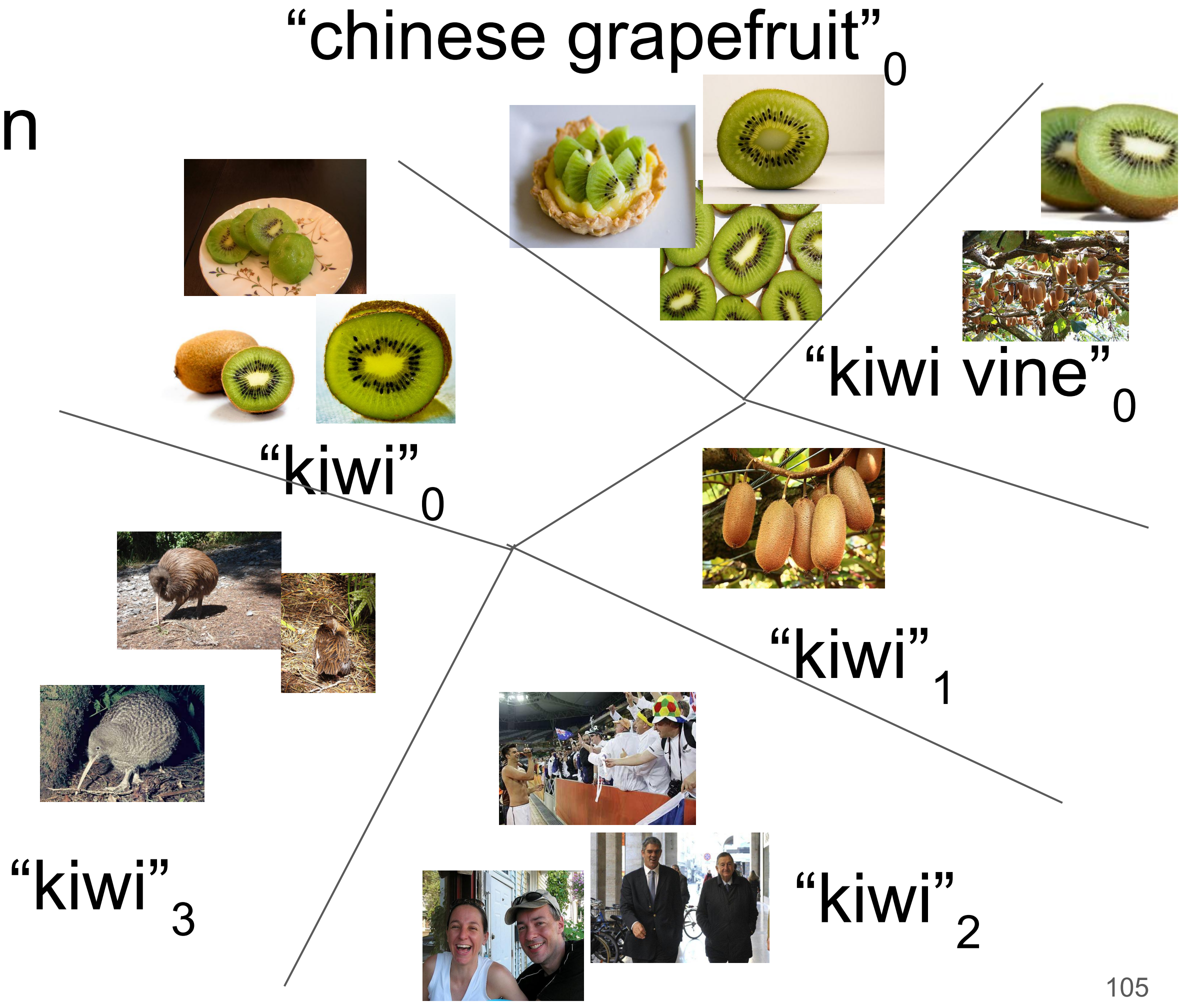
“kiwi”$_1$

“kiwi”$_2$

“kiwi”$_3$

Synonymy Detection

- Greedily merge nearest neighboring clusters
- Use cluster (sense) means to represent them
- Cap merge maximum senses (20, in our experiments)
- Results in synsets



“chinese grapefruit”$_0$

“kiwi vine”$_0$

“kiwi”$_0$

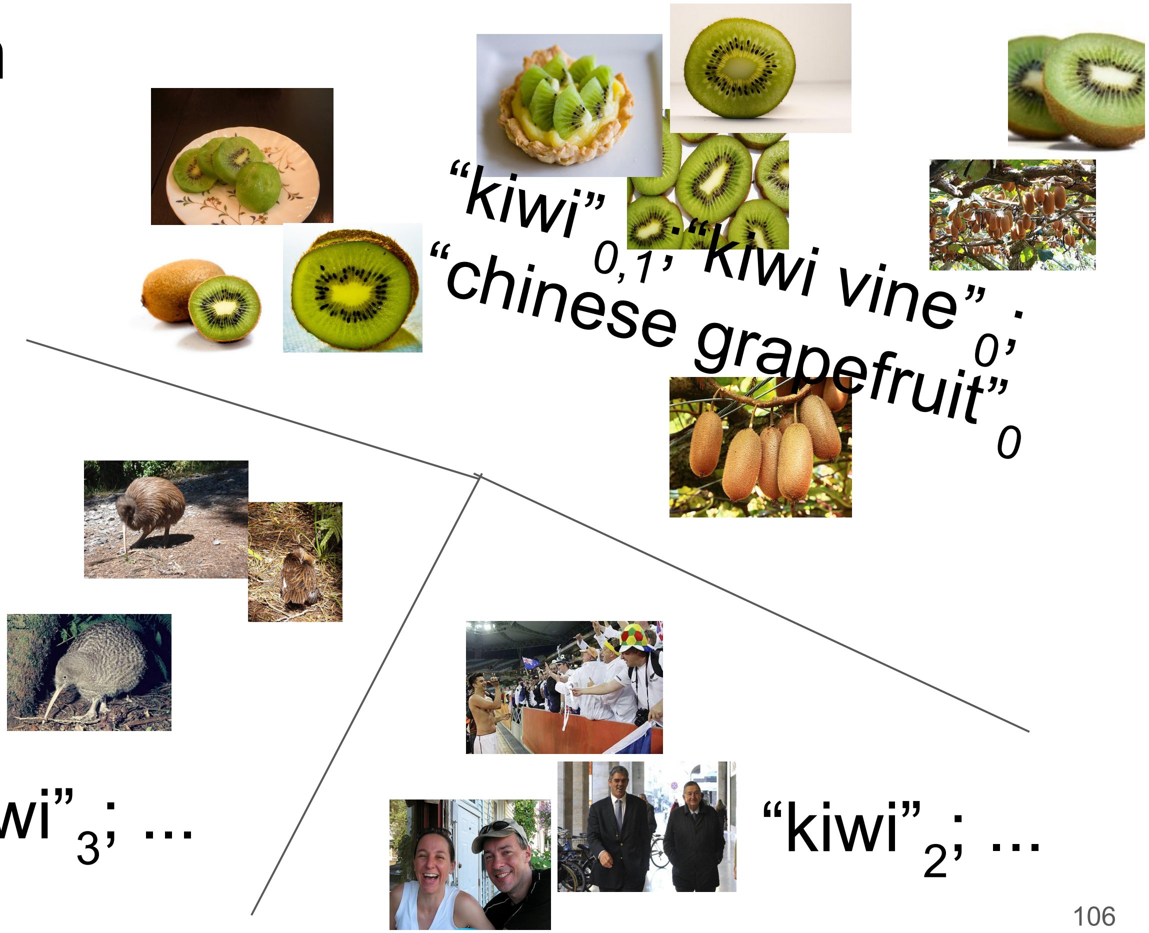
“kiwi”$_1$

“kiwi”$_3$

“kiwi”$_2$

Synonymy Detection

- Greedily merge nearest neighboring clusters
- Use cluster (sense) means to represent them
- Cap merge maximum senses (20, in our experiments)
- Results in synsets



“kiwi”$_{0,1}$; “kiwi vine”$_0$; “chinese grapefruit”$_0$

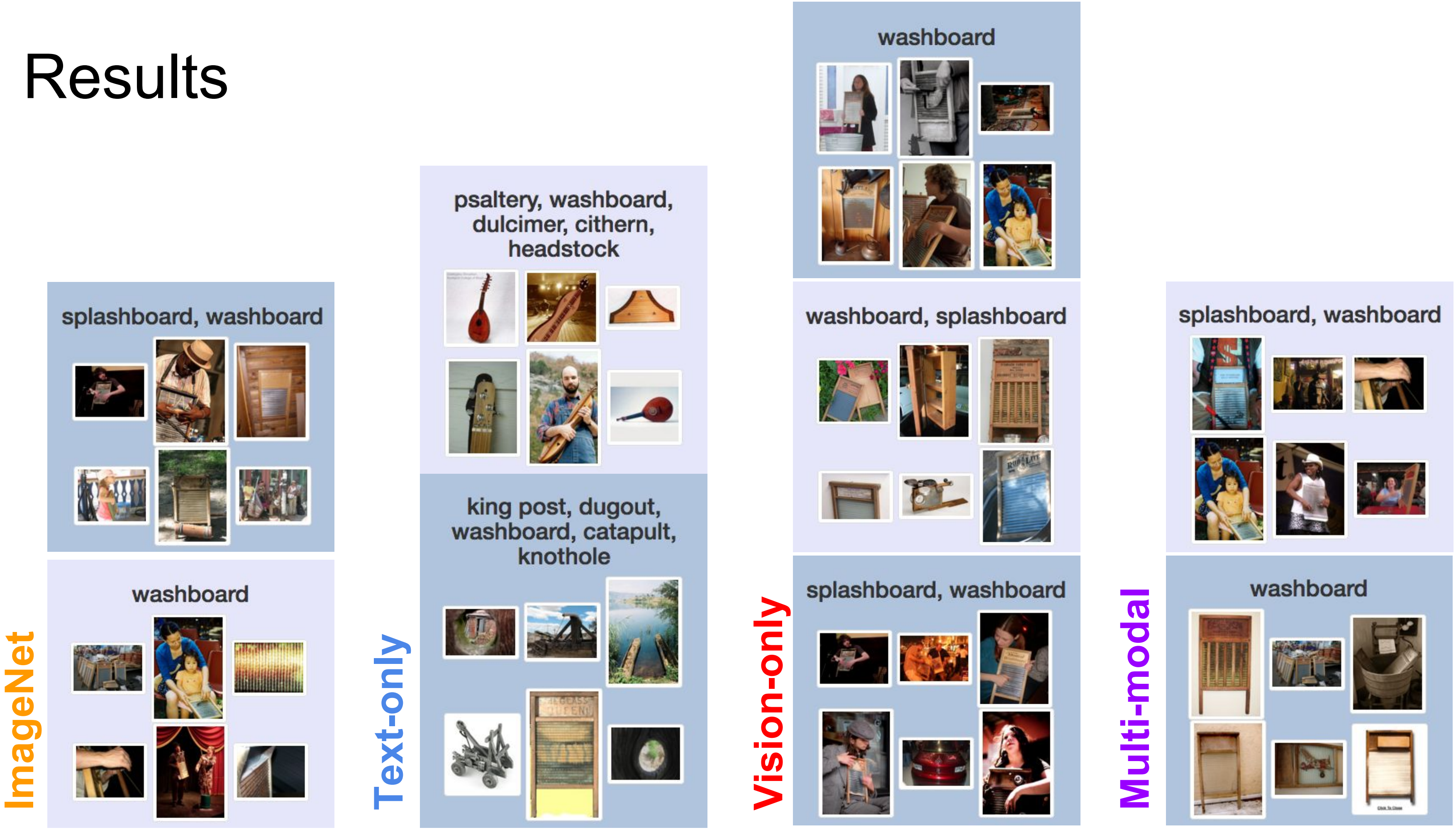
“kiwi”$_3$; ...

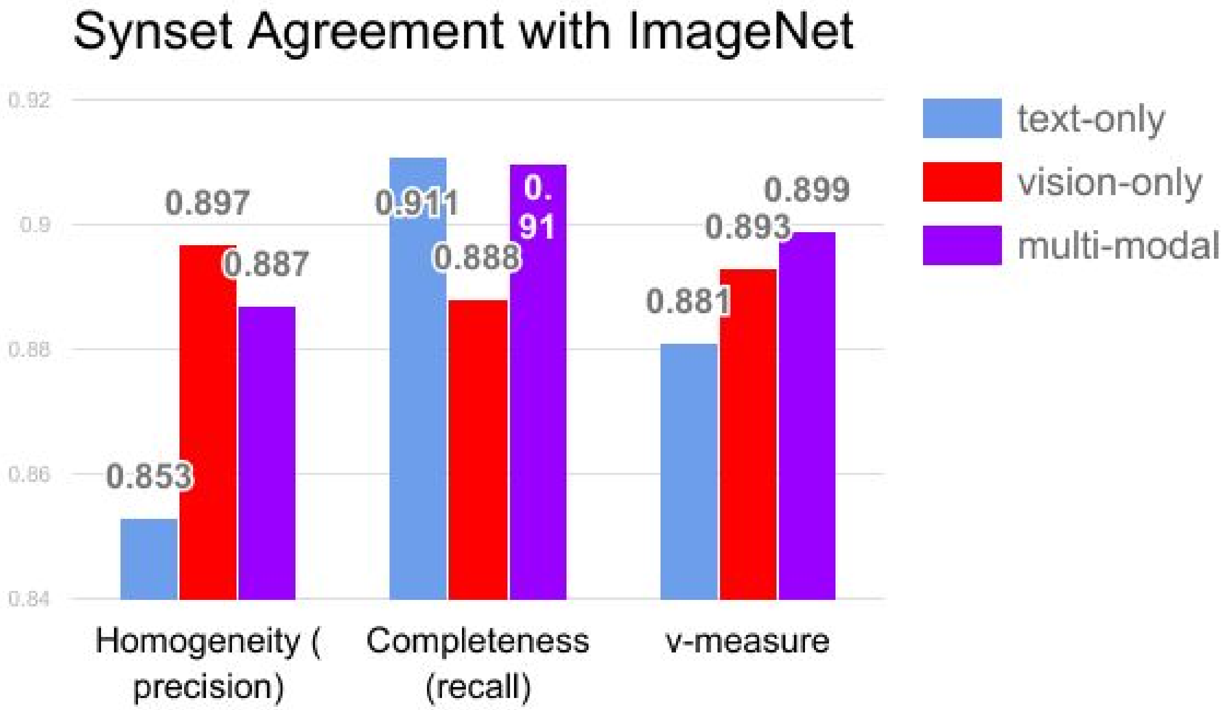
“kiwi”$_2$; ...

Experiments

- Held out the synsets used to train the VGG as validation data
- Set hyperparameters for clustering and induced LSA text feature space from validation data
- Ran word sense induction and synonymy detection with text-only, visual-only, and multi-modal features
- Measure homogeneity, completeness, and their harmonic mean between induced synsets and ImageNet synsets
    - Analogous to precision, recall, and *f*-measure for sets of sets [Manandhar et al., 2010],

Results

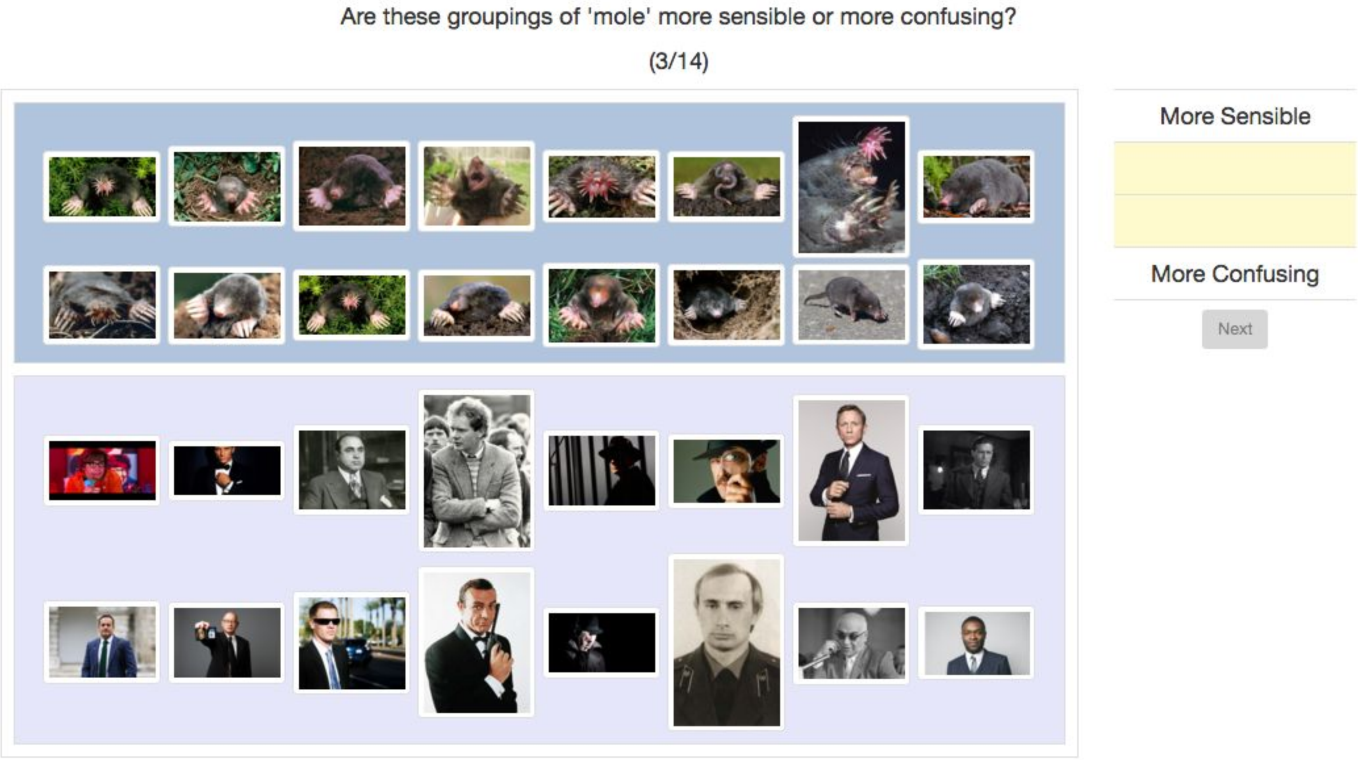

ImageNet
splashboard, washboard
washboard
Text-only
psaltery, washboard, dulcimer, cithern, headstock
king post, dugout, washboard, catapult, knothole
Vision-only
washboard
washboard, splashboard
splashboard, washboard
Multi-modal
splashboard, washboard
washboard


Synset Agreement with ImageNet

Human Evaluations

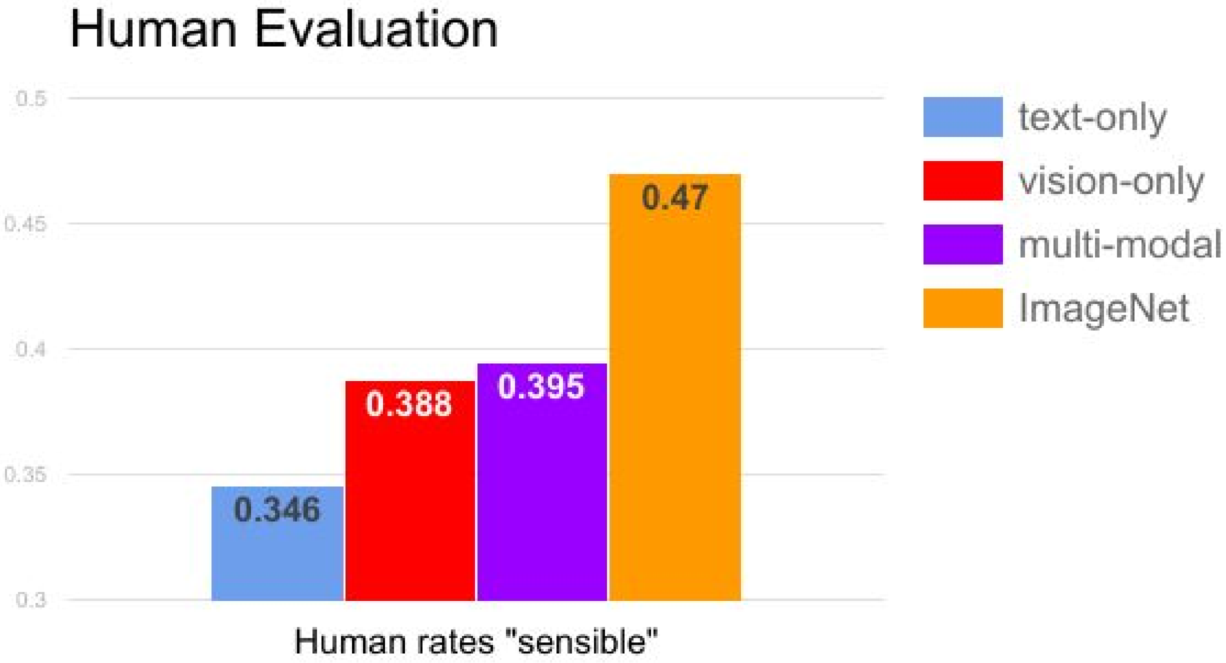
- Synset induction tends to join things ImageNet separates
- ImageNet separates people by nationality (e.g. “Austrian” and “Croatian”)
- ImageNet has odd categories for describing people (e.g. “energizer”)
- We evaluate induced synsets and ImageNet synsets by human judgements of sensibility
    - Humans shown all synsets a sampled noun phrase ended up in for each system
- Use paired t-test to determine whether humans statistically significantly favor ImageNet over induced synsets

Human Evaluations



Are these groupings of 'mole' more sensible or more confusing?

(3/14)



- Text-only and vision-only statistically significantly less favored versus ImageNet
- Multi-modal difference not significant

Findings

- Synset induction can be used to create ImageNet-like resource at leaf level from observations tagged with single labels
- Image and text features together lead to synsets that more closely match ImageNet's
- Human annotators rate multi-modal synsets sensible 84% as often as ImageNet synsets

Natural Language Understanding for Robots

> Go to Alice’s office and
> get the light mug for
> the chair.

- Commands that need to be actualized through robot action
- World knowledge about people and the surrounding office space
- Perception information to identify referent object
  - Now we have methodology to identify senses of “light” ✓

Natural Language Understanding for Robots

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

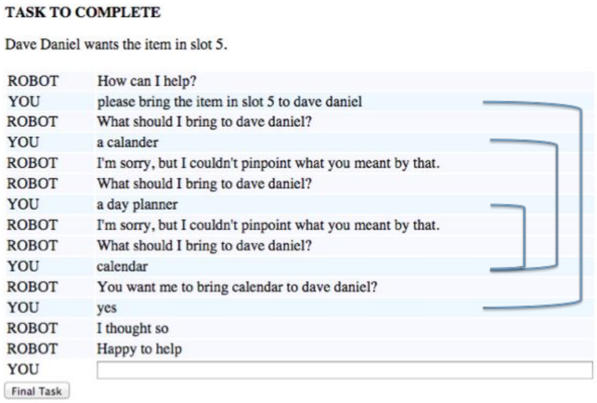
Natural Language Understanding for Robots

> Go to Alice’s office and
> get the light mug for
> the chair.

- Commands that need to be actualized through robot action
- World knowledge about people and the surrounding office space
- Perception information to identify referent object

Natural Language Understanding for Robots

Go to Alice’s office and get the light mug for the chair.



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	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	

- Commands that need to be actualized through robot action ✓
- World knowledge about people and the surrounding office space ✓
- Perception information to identify referent object

Natural Language Understanding for Robots

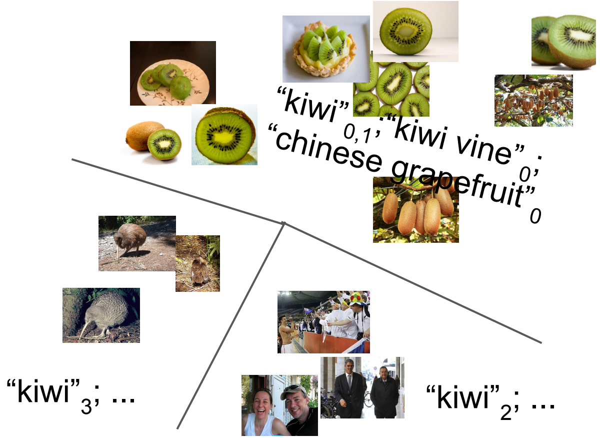
Go to Alice's office and get the light mug for the chair.

Human Turn
The description offered by the subject provides positive labels for chosen object.

- Commands that need to be actualized through robot action ✓
- World knowledge about people and the surrounding office space ✓
- Perception information to identify referent object ✓

Natural Language Understanding for Robots

Go to Alice’s office and get the light mug for the chair.



- Commands that need to be actualized through robot action ✓
- World knowledge about people and the surrounding office space ✓
- Perception information to identify referent object ✓
    - With methods to handle polysemy and synonymy ✓

Future Directions

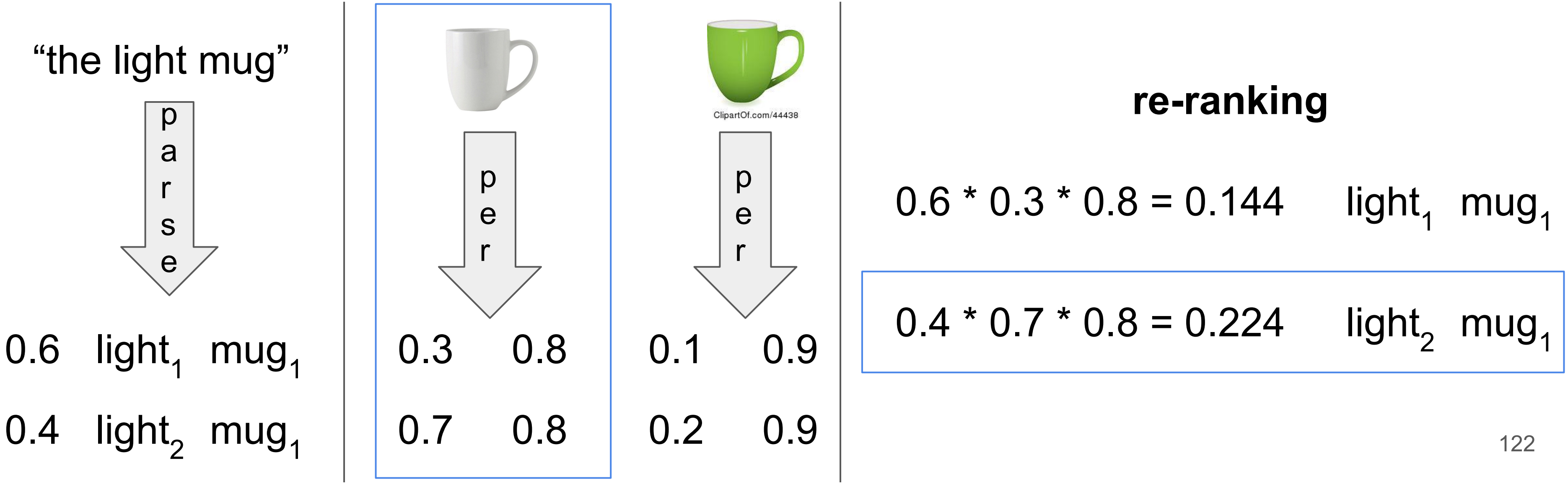
- Synset induction for multi-modal, perceptually grounded predicates
- Grounding semantic parses against both knowledge and perception
- New opportunities for continuous learning

Synset Induction for Grounded Predicates

- Differs from completed work on synset induction
- Multiple labels per object, rather than single noun phrase associated with each
- Completed work with two modalities simply averaged representation vector distances
- With many multiple perceptual contexts, more sophisticated combination strategies may be possible
    - For example, “light” senses are visible by comparing context relevance

Semantic Re-ranking from Perception Confidence

- Parser can return many parses, ranked with confidence values
- Perception predicates return confidence per object in the environment
- Combine confidences to get joint decision on understanding



“the light mug”

ClipartOf.com/44438

0.6 light$_1$ mug$_1$

0.4 light$_2$ mug$_1$

0.3	0.8
0.7	0.8

0.1	0.9
0.2	0.9

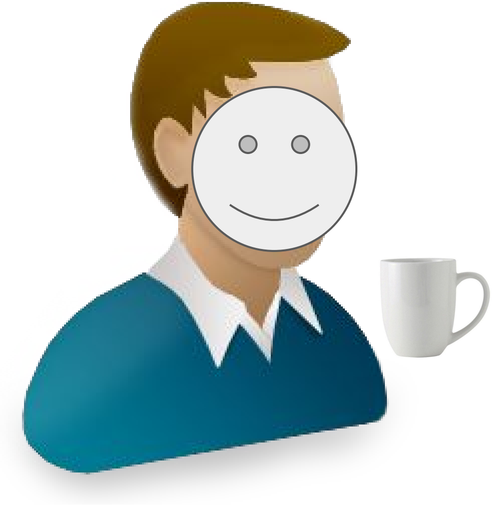
re-ranking

0.6 * 0.3 * 0.8 = 0.144 light$_1$ mug$_1$

0.4 * 0.7 * 0.8 = 0.224 light$_2$ mug$_1$

Perception Training Data from Dialog

- “Bring me the light mug”
- Human can confirm correct object was delivered
- Then delivered object is positive example for $\textit{light}_2$ and $\textit{mug}_1$



Natural Language Understanding for Robots

Go to Alice’s office and get the light mug for the chair.

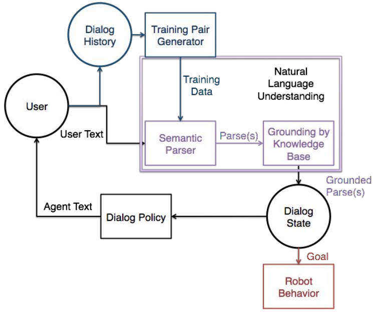
Natural Language Understanding for Robots

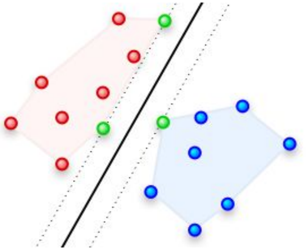
I will go to Room 1, pick up a light mug object, and deliver it to Bob.

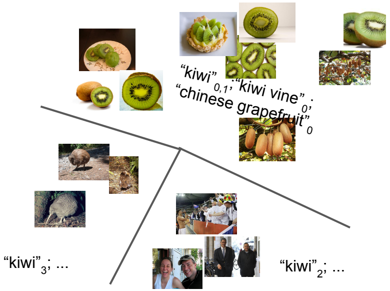
Natural Language Understanding for Robots

“alert me if her heart rate decreases”
“bring me his chart”
“go and get the family”
“scalpel”

“text me when the speaker arrives”
“grab the heavy, green mug”
“lead him to alice’s office”
“get out of the way”



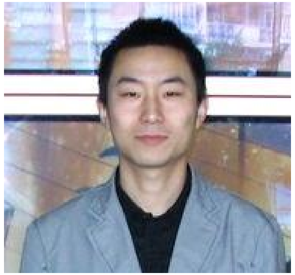




Thanks!

- Dissertation Committee

- UTCS Robotics Collaborators



- NSF, Stefanie Tellex, Brown University Computer Science, and you

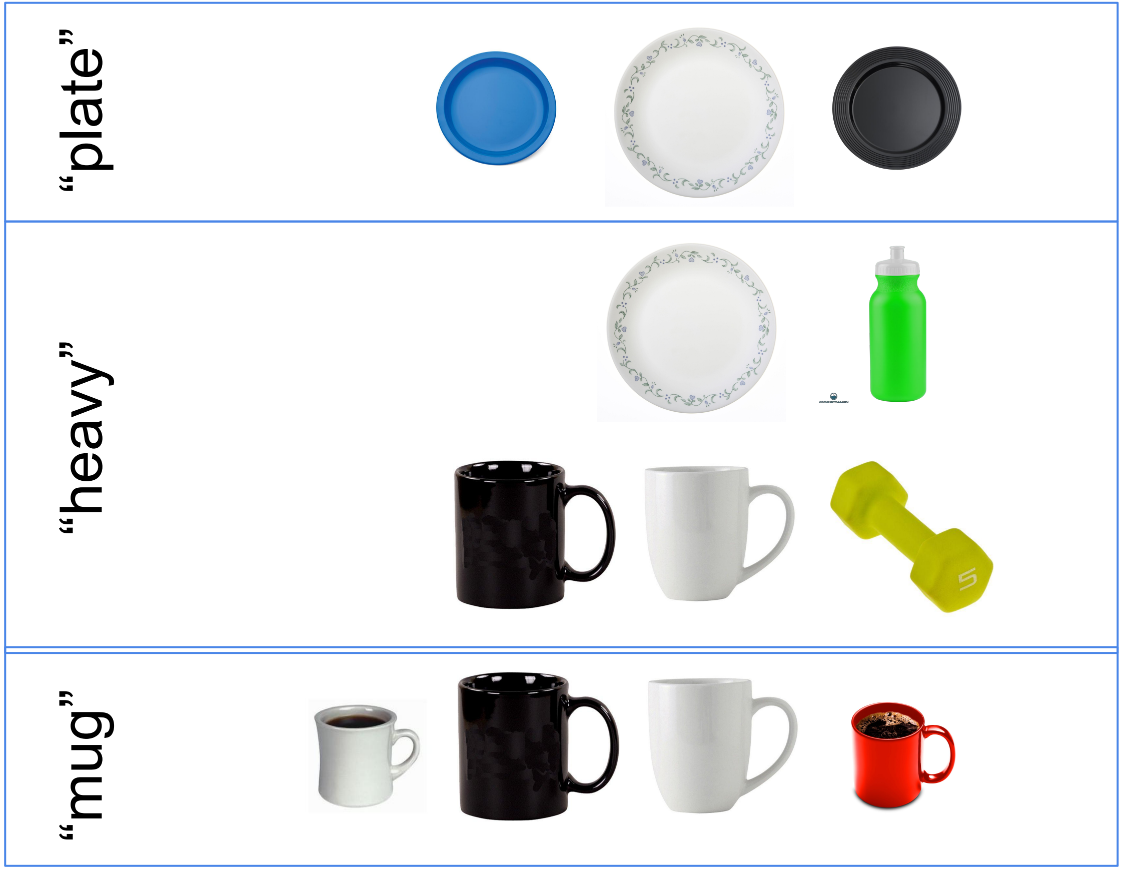
Continuously Improving Robotic Natural Language Understanding with Semantic Parsing, Dialog, and Multi-modal Perception

Jesse Thomason
University of Texas at Austin

Graded Adjectives

- Think of gradation as a form of polysemy
- Semantic parser can use surrounding context
- Re-ranking of parses, as discussed, can help disambiguate

words

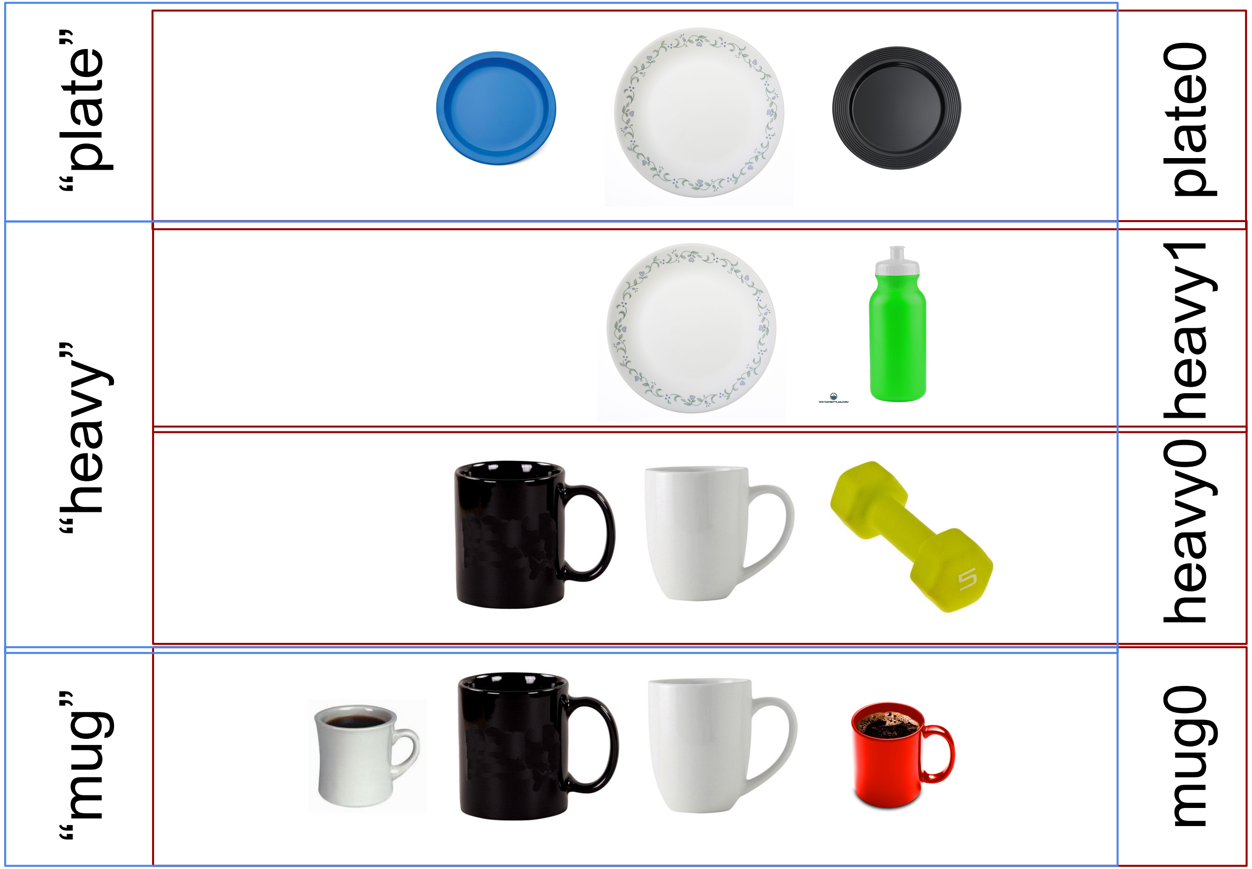


“plate”

“heavy”

“mug”

words



"plate"

"heavy"

"mug"

plate0

heavy1

heavy0

mug0

predicates

Comparative Adjectives

- E.g. "taller”, “heavier”
- Take two arguments: obj1, obj2
- Train classifier on the feature differences between obj1, obj2
- Can otherwise be handled with existing architecture
- Superlatives: majority winner object in pairwise comparative

Sparse Perceptual Data

Human Turn

The description offered by the subject provides positive labels for chosen object.

“An empty metallic aluminum container”

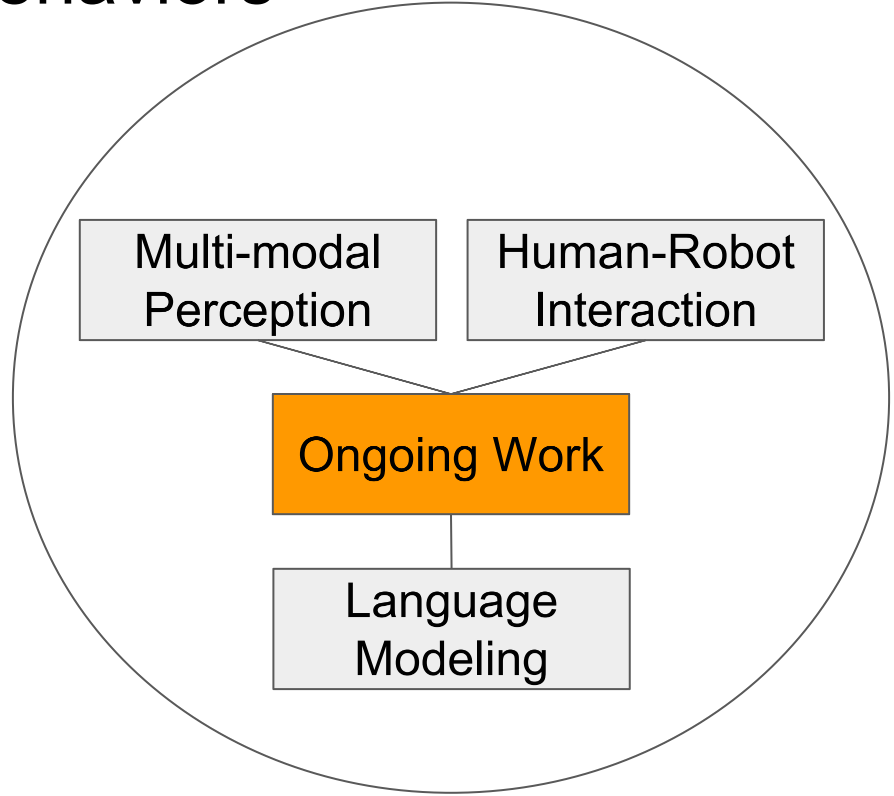
+ negative examples from follow-up questions

Sparse Perceptual Data

Kappas with human labels using per-context-xval distributed per predicate					
	drop/audio	**drop/haptic**	**look/color**	**...**	**press/haptic**
red	.057	.065	.074	...	.051
half-full	.072	.064	.017	...	.063
...	...	...	...	...	...
aluminum	.10	.075	.075	...	.055

- Spurious co-occurrences give misleading kappas
    - What if your sparse sample of yellow objects are all heavy?

Guiding Language Grounding with Multiple Interaction Behaviors



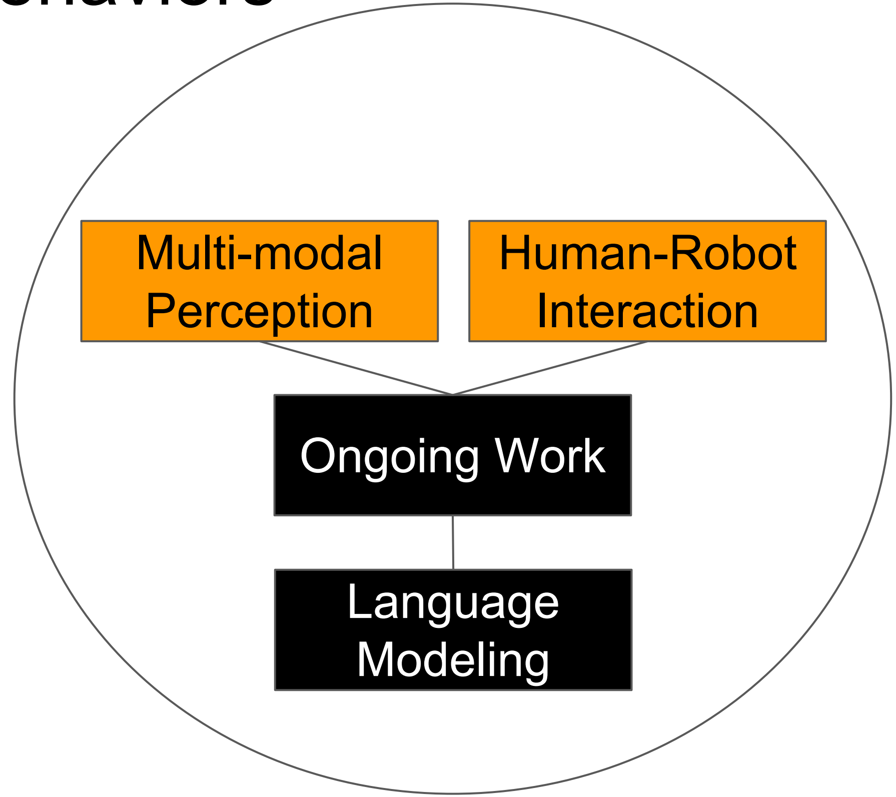
Multi-modal Perception

Human-Robot Interaction

Ongoing Work

Language Modeling

Guiding Language Grounding with Multiple Interaction Behaviors



Room for More Information from Humans

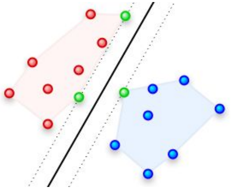
Human Turn

The description offered by the subject provides positive labels for chosen object.

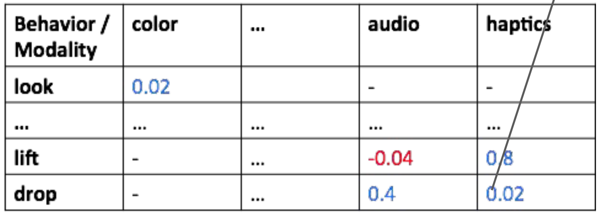
Human:“An empty metallic aluminum container”

Robot: “Would you use the word “empty” to describe this object?”

Sensorimotor context SVM



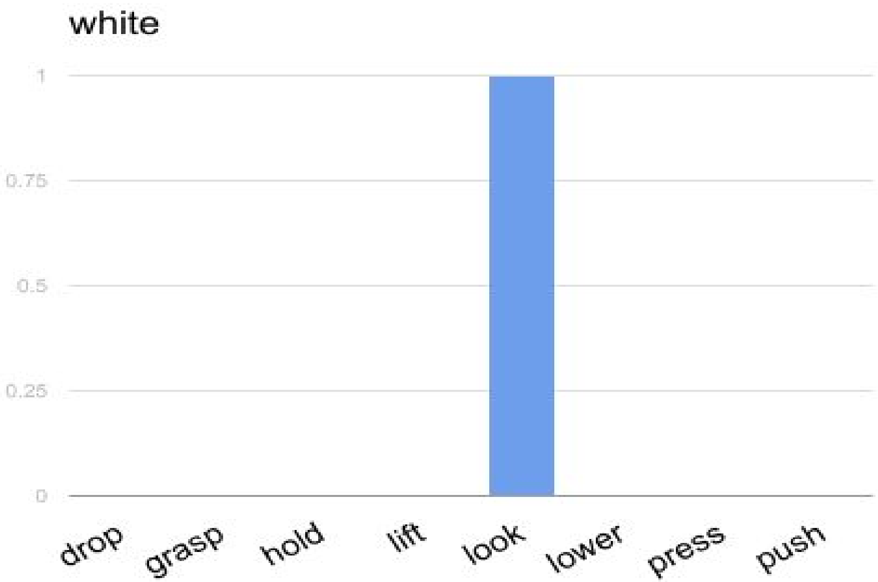
“empty”?

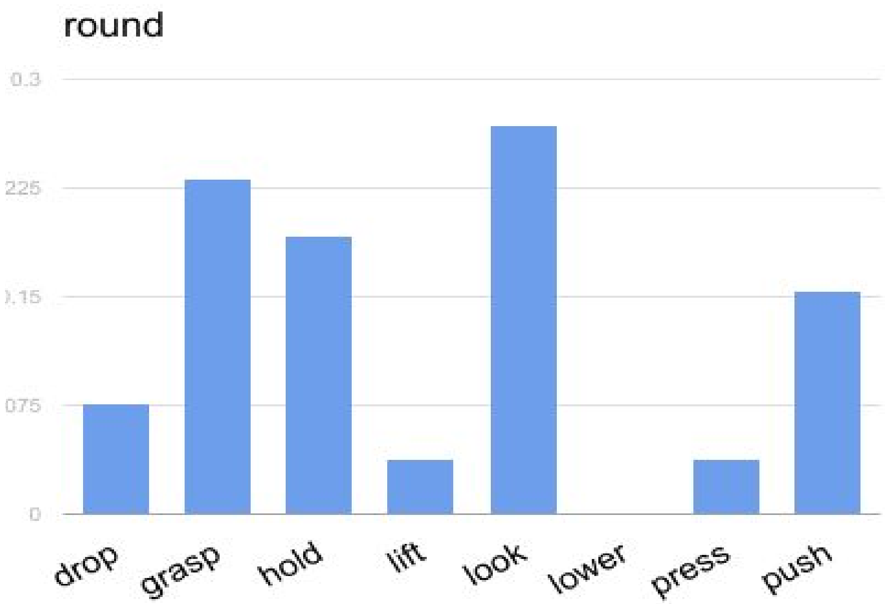


Behavior / Modality	color	...	audio	haptics
look	0.02		-	-
...	...	...	...	...
lift	-	...	-0.04	0.8
drop	-	...	0.4	0.02

Robot: “How can you tell if something can be described as “empty”?”

Human: “You can pick it up.”


white
1
0.75
0.5
0.25
0
drop
grasp
hold
lift
look
lower
press
push

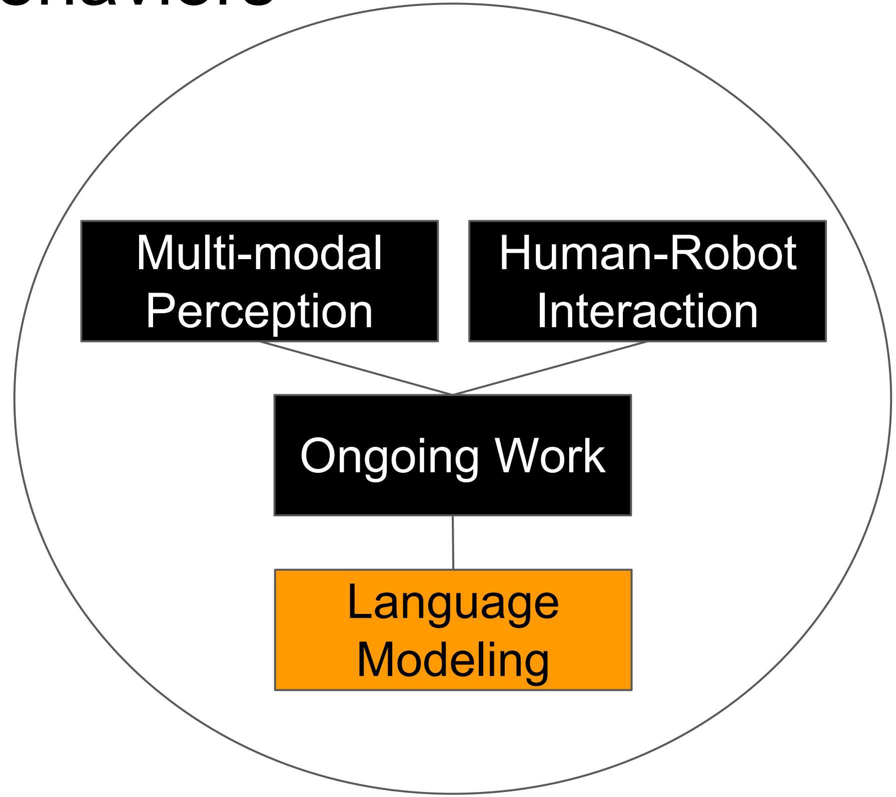

round
0.3
225
).15
075
0
drop
grasp
hold
lift
look
lower
press
push

Room for More Information from Humans

Kappas with human labels using per-context-xval distributed per predicate					
	grasp/audio	**grasp/haptic**	**look/color**	**...**	**lift/haptic**
red	0	0	.5	...	0
half-full	0	0	0	...	.25
...	...	...	...	...	...
aluminum	.25	.25	.25	...	0

- Use human annotations to restrict contexts to relevant behaviors
  - Makes spurious kappas less likely by masking irrelevant behaviors

Guiding Language Grounding with Multiple Interaction Behaviors



Multi-modal Perception

Human-Robot Interaction

Ongoing Work

Language Modeling

Room for More Information from Language

- In past work, decision is made for each predicate *p* on object *o* as

$$d = \sum_{c \text{ in contexts}} \kappa_{c,p} \mathrm{SVM}_c(o)$$

- With the sign of d determining whether *p* applies (each SVM returns 1 or -1)
- Thus, for each context *c*, we consider only the confidence kappa associated with predicate *p*
- Intuition: if predicate *q* is similar to predicate *p* and has high confidence in context *c*, maybe *p* should too
    - “Green” is similar to “mauve”, so maybe we should trust look/color for mauve too

Room for More Information from Language

- Calculate the cosine similarity between every predicate pair in word2vec space and set confidence based on kappas from similar predicates
  - Our cosine similarity ranges in [0, 1] with distances less than 0 rounded up
- Then the decision for predicate *p* with embedding becomes

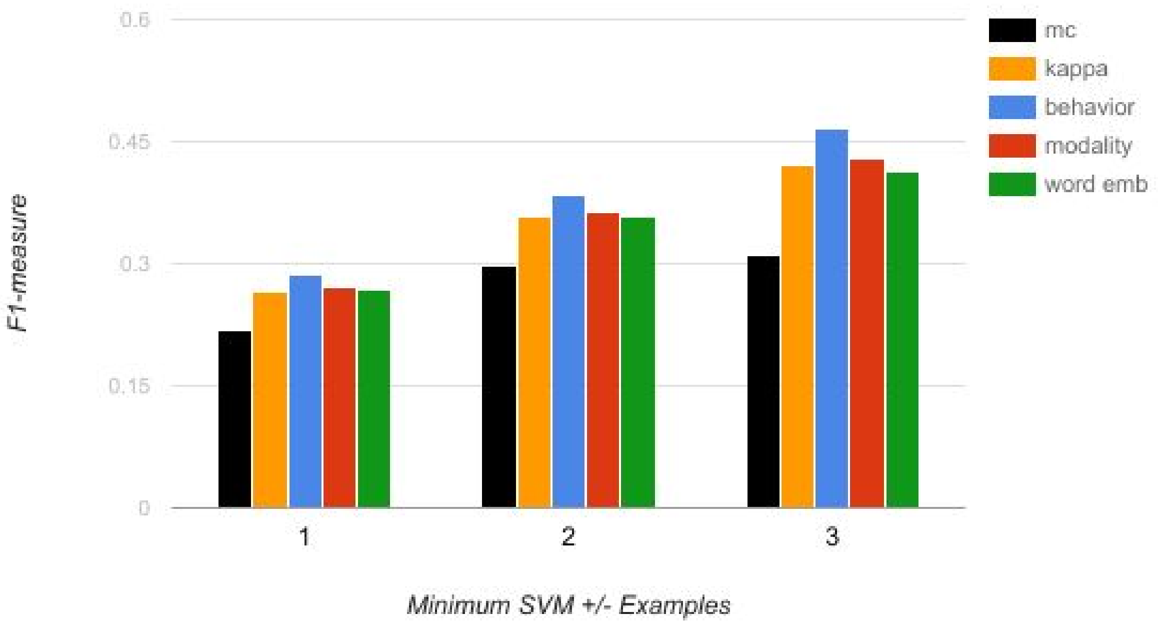
$$\text{we}\kappa_{c,p} = \sum_{q \text{ in predicates}} \kappa_{c,q} \cos(e_p, e_q)$$

$$d = \sum_{c \text{ in contexts}} \text{we}\kappa_{c,p} \text{SVM}_c(o)$$

Experiments

- Gather annotations for behaviors after demonstrating them on a sample object
- “What behaviors would you engage in to determine if ____ could be used to describe the object?”
- Six of 14 annotators used, with average kappa=0.47 (moderate agreement)
- We use Google News embeddings to embed our predicates, getting cosine similarities for 76 out of 81 of them
    - Missing words are hyphenated like “half-full” or odd compounds like “spraycan”
    - Missing words given uniform distance to one another

(Preliminary) Results



- Adding behavior and modality annotations helps
- Adding word embeddings may generalize meanings too much

Findings

- Going beyond obtaining true/false labels on a per predicate basis for objects may speed perceptual grounding with sparse data
- Potential to reduce exploratory behaviors needed on a new object
  - To determine if something is “green”, we only need to look at it
- Adding unsupervised information from large text corpora allows us to share label information
  - Lots of labels for “green” and few for “mauve” but we know “mauve” is a color and can avoid spurious results from other contexts