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

Jesse Thomason
Doctoral Dissertation Proposal
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 and use natural language from humans
Go to Alice’s office and get the light mug for the chair.
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 with given robot and domain
- Interaction with humans should strengthen understanding over time
Outline

● Background
● Completed work
● Proposed Work
● Conclusion
Background: Situating this Proposal

Semantic Parsing

This proposal

Language Grounding
Background: Situating this Proposal

Semantic Parsing

- Commanding Robots
- Dialog
- Semantic Understanding
  - Thomason, 2015

Language Grounding

- Multi-modal Perception
- Grounding
  - Thomason, 2016
- Human-robot Interaction
Background: Situating this Proposal

Semantic Parsing
Thomason, 2015

Language Grounding
- Word-sense Induction
- Multi-modal Perception
- Grounding
  - Synonymy Detection
  - Human-robot Interaction

Thomason, in progress
Thomason, 2016
Background: Situating this Proposal

This proposal

Semantic Parsing
Thomason, 2015

Language Grounding
Thomason, 2016
Thomason, in progress
Outline

● Background
  ○ Semantic Parsing
  ○ Language Grounding
Go to Alice’s office and get the light mug for the chair.

go(\(\lambda x.\) (office(\(x\)) \land owns(alice, \(x\)))) \land 
deliver(\(\lambda y.\) (light2(\(y\)) \land mug1_cup2(y))), 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

<table>
<thead>
<tr>
<th>NP</th>
<th>NP\NP/N</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>'s</td>
<td>office</td>
</tr>
</tbody>
</table>
Background: Semantic Parsing

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

- Leaf-level semantic meanings can be propagated through tree

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

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

\[
\lambda P. \lambda y. (\text{the}(\lambda x. (P(x) \land \text{owns}(y, x))))
\]

Alice

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

\[
\begin{align*}
goose((\forall x. (\text{office}(x) \land \text{owns}(\text{alice}, x)))) & \land \\
\text{deliver}((\forall y. (\text{light2}(y) \land \text{mug1_cup2}(y))), \text{bob})
\end{align*}
\]
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 features
  - CCG combination features
  - Word -> semantics features
- 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
Outline

● Background
  ○ Semantic Parsing
  ○ Language Grounding
Background: Language Grounding

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

- World knowledge about people and the surrounding office space
- Perception information to identify referent object
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”

\[
\text{the}(\lambda x. (\text{office}(x) \land \text{owns}(\text{alice}, x)))
\]
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"

\[ \text{the}(\exists y. (\text{light2}(y) \land \text{mug1_cup2}(y))) \]
Background: Language Grounding

**word**

**instances**

“light”

“mug”

“cup”
## Background: Language Grounding

<table>
<thead>
<tr>
<th>word</th>
<th>“light”</th>
<th>“mug”</th>
<th>“cup”</th>
</tr>
</thead>
<tbody>
<tr>
<td>instances</td>
<td>light1, light2</td>
<td>mug1, cup2</td>
<td>cup1</td>
</tr>
</tbody>
</table>
Outline

● Background

● Completed work
  ○ Learning to Interpret Natural Language Commands through Human-Robot Dialog
  ○ Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”
  ○ Multi-modal Word Synset Induction

● Proposed Work

● Conclusion
Learning to Interpret Natural Language Commands through Human-Robot Dialog

Semantic Parsing

Commanding Robots  Dialog

Thomason, 2015

Semantic Understanding
**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

**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:**

1. Coffee
2. Sandwich
3. Calendar
4. Trash can
5. Planner
Semantic Parsing

Commanding Robots

Dialog

Thomason, 2015

Semantic Understanding
Dialog

- User
  - User Text
  - Dialog Policy
    - Determines how system responds given the state
- Natural Language Understanding
  - Abstract Representation of User Meaning
  - Dialog State
    - Represents current beliefs of the system given dialog history

Dialog + Commanding Robots

- Past work uses dialog as part of a pipeline for commanding robots (Matuszek, 2012; Mohan, 2012)
- Adding a dialog component allows the robot to refine its understanding
Dialog + Commanding Robots
Semantic Parsing

Commanding Robots

Dialog

Thomason, 2015

Semantic Understanding
Past work uses semantic parsing as an understanding step to command robots (Kollár, 2013)
Semantic Parsing

Commanding Robots

Dialogue

Thomason, 2015

Semantic Understanding
Generating New Training Examples

- Past work generates training data for a parser given a corpus of conversations (Artzi, 2011)
- We pair confirmed understanding from dialog with previous misunderstandings
Generating New Training Examples

<table>
<thead>
<tr>
<th>TASK TO COMPLETE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dave Daniel wants the item in slot 5.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ROBOT</th>
<th>How can I help?</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOU</td>
<td>please bring the item in slot 5 to dave daniel</td>
</tr>
<tr>
<td>ROBOT</td>
<td>What should I bring to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>a calander</td>
</tr>
<tr>
<td>ROBOT</td>
<td>I'm sorry, but I couldn't pinpoint what you meant by that.</td>
</tr>
<tr>
<td>ROBOT</td>
<td>What should I bring to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>a day planner</td>
</tr>
<tr>
<td>ROBOT</td>
<td>I'm sorry, but I couldn't pinpoint what you meant by that.</td>
</tr>
<tr>
<td>ROBOT</td>
<td>What should I bring to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>calendar</td>
</tr>
<tr>
<td>ROBOT</td>
<td>You want me to bring calendar to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>yes</td>
</tr>
<tr>
<td>ROBOT</td>
<td>I thought so</td>
</tr>
<tr>
<td>ROBOT</td>
<td>Happy to help</td>
</tr>
<tr>
<td>YOU</td>
<td></td>
</tr>
</tbody>
</table>

Final Task
Generating New Training Examples

**TASK TO COMPLETE**

Dave Daniel wants the item in slot 5.

<table>
<thead>
<tr>
<th>ROBOT</th>
<th>How can I help?</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOU</td>
<td>please bring the item in slot 5 to dave daniel</td>
</tr>
<tr>
<td>ROBOT</td>
<td>What should I bring to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>a calander</td>
</tr>
<tr>
<td>ROBOT</td>
<td>I'm sorry, but I couldn't pinpoint what you meant by that.</td>
</tr>
<tr>
<td>ROBOT</td>
<td>What should I bring to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>a day planner</td>
</tr>
<tr>
<td>ROBOT</td>
<td>I'm sorry, but I couldn't pinpoint what you meant by that.</td>
</tr>
<tr>
<td>ROBOT</td>
<td>What should I bring to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>calendar</td>
</tr>
<tr>
<td>ROBOT</td>
<td>You want me to bring calendar to dave daniel?</td>
</tr>
<tr>
<td>YOU</td>
<td>yes</td>
</tr>
<tr>
<td>ROBOT</td>
<td>I thought so</td>
</tr>
<tr>
<td>ROBOT</td>
<td>Happy to help</td>
</tr>
<tr>
<td>YOU</td>
<td></td>
</tr>
</tbody>
</table>

*Final Task*
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

[Final Task]
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**

*Final Task*
### Generating New Training Examples

<table>
<thead>
<tr>
<th>Utterance</th>
<th>please bring the item in slot 5 to dave daniel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meaning</strong></td>
<td>$action(\text{bring}) \land \text{patient(\text{bring, calendar})}$ $\land \text{recipient(\text{bring, dave})}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Utterance</th>
<th>a calander</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meaning</strong></td>
<td>$\text{calendar}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Utterance</th>
<th>a day planner</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Meaning</strong></td>
<td>$\text{calendar}$</td>
</tr>
</tbody>
</table>
Generating New Training Examples

<table>
<thead>
<tr>
<th>Utterance</th>
<th>please <em>bring</em> the <em>item in slot 5</em> to <em>dave daniel</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Meaning</td>
<td>$action(\text{bring}) \land patient(\text{bring, calendar}) \land recipient(\text{bring, dave})$</td>
</tr>
<tr>
<td>Utterance</td>
<td><em>a calander</em></td>
</tr>
<tr>
<td>Meaning</td>
<td><em>calendar</em></td>
</tr>
<tr>
<td>Utterance</td>
<td><em>a day planner</em></td>
</tr>
<tr>
<td>Meaning</td>
<td><em>calendar</em></td>
</tr>
</tbody>
</table>
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
**TASK TO COMPLETE**

Dave Daniel wants the item in slot 5.

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

**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:**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Mechanical Turk Experiment

- Four batches of ~100 users each
- Retraining after every batch (~50 training goals)
- Performance measured every batch (~50 testing goals)
Mechanical Turk Dialog Turns

### Navigation task
- average Turker Turns for success

<table>
<thead>
<tr>
<th>Batch</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.7</td>
<td>2.5</td>
<td>3.4</td>
<td>3.5</td>
</tr>
</tbody>
</table>

### Delivery task
- average Turker turns for success

<table>
<thead>
<tr>
<th>Batch</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18</td>
<td>5.6</td>
<td>5.2</td>
<td>3.8</td>
</tr>
</tbody>
</table>
Mechanical Turk Survey Responses

The robot understood me

- Strongly Agree
  - Batch 0: 2.2
  - Batch 1: 2.6
  - Batch 2: 2.7
  - Batch 3: 2.8

- Somewhat Agree
- Average
- Somewhat Disagree
- Strongly Disagree
Mechanical Turk Survey Responses

The robot frustrated me

<table>
<thead>
<tr>
<th></th>
<th>Batch 0</th>
<th>Batch 1</th>
<th>Batch 2</th>
<th>Batch 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly Agree</td>
<td>2.2</td>
<td>2</td>
<td>1.9</td>
<td>1.5</td>
</tr>
<tr>
<td>Somewhat Agree</td>
<td>2</td>
<td>2</td>
<td>1.9</td>
<td>1.5</td>
</tr>
<tr>
<td>Somewhat Disagree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strongly Disagree</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

**Navigation task completion rate**

- Batch 0: 90%
- Batch 1: 90%

**Delivery task Completion rate**

- Batch 0: 20%
- Batch 1: 60%
Segbot Survey Responses

The robot understood me

- **Strongly Agree**: Batch 0 = 1.6, Batch 1 = 2.9
- **Somewhat Agree**: Batch 0 = 1.6, Batch 1 = 2.9
- **Somewhat Disagree**: Batch 0 = 1.6, Batch 1 = 2.9
- **Strongly Disagree**: Batch 0 = 1.6, Batch 1 = 2.9
Segbot Survey Responses

The robot frustrated me

- Batch 0: Strongly Agree - 2.5
- Batch 1: Somewhat Agree - 1.5

Average scores for batch 0 and batch 1.
Contributions

- 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 use changes or new users with new lexical choices arrive, can adapt on-the-fly
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
Outline

● Background

● Completed work
  ○ Learning to Interpret Natural Language Commands through Human-Robot Dialog
  ○ Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”
  ○ Multi-modal Word Synset Induction

● Proposed Work

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

Language Grounding

Multi-modal Perception

Grounding

Thomason, 2016

Human-robot Interaction
"An empty metallic aluminum container"
Robot makes guesses until human confirms it found the right object.
Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”

Language Grounding

Multi-modal Perception

Grounding

Thomason, 2016

Human-robot Interaction
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
- 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”

Language Grounding

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, proprioceptive information, and haptic information (from arm motors)
- “Look” is just one way to explore; gather visual features such as VGG penultimate layer
- Feature representation of each object has many sensorimotor contexts
- Context is a combination of an exploratory behavior and associated sensory modality
Multi-Modal Perception

Grasp

Lift

Lower

Drop

Press

Push
Multi-Modal Perception

- Still need language labels for objects
- Annotating each object with every possible descriptor is unrealistic and boring
- 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)
- 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
- After human guesses correct, ask for explicit yes/no on whether some predicates apply to target
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 contribution by confidence using leave-on-out x-val over objects
Building Perceptual Classifiers

Sensorimotor context SVMs

Empty?

<table>
<thead>
<tr>
<th>Behavior / Modality</th>
<th>color</th>
<th>...</th>
<th>audio</th>
<th>haptics</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td>0.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>lift</td>
<td>-</td>
<td>...</td>
<td>-0.04</td>
<td>0.8</td>
</tr>
<tr>
<td>drop</td>
<td>-</td>
<td>...</td>
<td>0.4</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Decision gives sign

Kappa with human labels gives magnitude
Building Perceptual Classifiers

Empty?

$$0.02 + \ldots + (-0.04) + 0.8 + 0.4 + 0.02 = 1.37$$

<table>
<thead>
<tr>
<th>Behavior / Modality</th>
<th>color</th>
<th>...</th>
<th>audio</th>
<th>haptics</th>
</tr>
</thead>
<tbody>
<tr>
<td>look</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lift</td>
<td>-</td>
<td>...</td>
<td>-0.04</td>
<td>0.8</td>
</tr>
<tr>
<td>drop</td>
<td>-</td>
<td>...</td>
<td>0.4</td>
<td>0.02</td>
</tr>
</tbody>
</table>
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 Predicate Agreement

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

<table>
<thead>
<tr>
<th>Metric</th>
<th>System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>vision only</td>
</tr>
<tr>
<td>precision</td>
<td>.250</td>
</tr>
<tr>
<td>recall</td>
<td>.179</td>
</tr>
<tr>
<td>$F_1$</td>
<td>.196</td>
</tr>
</tbody>
</table>

- *: 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 ($r = 0.549$)
“A tall blue cylindrical container”
Contributions

● We move beyond vision for grounding language predicates
● 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”
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...
Outline

● Background

● Completed work
  ○ Learning to Interpret Natural Language Commands through Human-Robot Dialog
  ○ Learning Multi-Modal Grounded Linguistic Semantics by Playing “I Spy”
  ○ Multi-modal Word Synset Induction

● Proposed Work

● Conclusion
Multi-modal Word Synset Induction

Language Grounding

Word-sense Induction

Multi-modal Perception

Grounding

Thomason, *in progress*

Synonymy Detection

Thomason, 2016

Human-robot Interaction
Multi-modal Word Synset Induction

- Words from “I, Spy” do not have a one-to-one mapping with perceptual predicates
- “Light” can mean lightweight or light in color (polysemy)
- “Claret” and “purple” refer to the same property (synonymy)
- Words have one or more senses
- A group of synonymous senses is called a synset (synonym sense set)
Multi-modal Word Synset Induction

Language Grounding

- Word-sense Induction
- Multi-modal Perception
- Grounding

- Thomason, *in progress*
- Thomason, 2016
- Synonymy Detection
- Human-robot Interaction
Word Sense Induction

- Task of discovering word senses
- “Bat”
  - Baseball, animal
- “Light”
  - Weight, color
- “Kiwi”
  - Fruit, bird, people
- Represent instances as vectors of their context; cluster to find senses
Multi-modal Word Synset Induction

Language Grounding

- Word-sense Induction
  - Thomason, *in progress*

- Multi-modal Perception
  - Thomason, 2016

- Synonymy Detection

- Grounding
  - Human-robot Interaction
Synonymy Detection

- Given words or word senses, find synonyms
- “Claret” and “purple”
- “Round” and “circular”
- “Kiwi” and “New Zealander” (some some sense of “kiwi”)
- Represent instances as vectors of their context; cluster means to find synonyms
Multi-modal Word Synset Induction

Language Grounding

Word-sense Induction

Thomason, in progress

Synonymy Detection

Multi-modal Perception

Thomason, 2016

Human-robot Interaction

Grounding
Multi-modal Perception

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

- Gather synsets and images from ImageNet
- All leaves; mix of polysemous, synonymous, and neither polysemous nor synonymous noun phrases
- Provides “gold” synsets we can aim to reconstruct from image-level instances
ImageNet Synsets to Mixed-sense Noun Phrases

- 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
Goal

- Reconstruct ImageNet-like synsets
- First perform word-sense induction on mixed-sense noun phrase inputs
- Given induced word senses, perform synonymy detection to form synsets
- Use reverse-image search to find webpages of text for each image
  - Get textual features and perform methods in multi-modal space
Word Sense Induction

Mixed-Sense Noun Phrase Data

Induced Word Senses
Synonymy Detection

Induced Word Senses

Reconstructed Synsets

kiwi sense 1

reconstructed synset 17
'kiwi', 'chinese gooseberry', 'kiwi vine'

chinese gooseberry sense 0

kiwi vine sense 0
Methods

- Commonly used VGG network to generate visual features (Simonyan 2014)
- Latent semantic analysis (LSA) of web pages to form textual feature space
- Images used to train VGG held out as development data for LSA and setting parameters
Methods

● Word sense induction
  ○ Use non-parametric $k$-means approach based on the gap statistic (Tibshirani 2001) to discover senses

● Synonymy detection
  ○ Use a nearest-neighbor method to join senses into synsets up to a pre-specified number of synsets estimated from development data
Preliminary Results

● Evaluate match of reconstructed and ImageNet synsets using $\nu$-measure (Rosenberg, 2007) and paired $f$-measure

● Quantitative evaluation unsurprising but disappointing
  ○ Precision-like metrics improved by polysemy detection (WSI)
  ○ Recall-like metrics improved by synonymy detection

● Multi-modal pipeline for both outperforms uni-modal pipelines

● ImageNet synsets are actually quite noisy and hard to recreate unsupervised
Preliminary Results

- ImageNet synsets are actually quite noisy and hard to recreate unsupervised
- “Austrian” and “Ukrainian” in separate synsets
- “Energizer” in a synset containing pictures of people in suits
- We plan a human evaluation to establish the better interpretability of our reconstructed synsets versus ImageNet’s
  - For example, our methods construct big synsets full of people for noun phrases “Austrian”, “Ukrainian”, “kiwi”, “energizer”, etc
Outline

- Background
- Completed work
- Proposed Work
- Conclusion
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

- Our proposed work focuses on integrating completed work to accomplish all these understanding components at once.

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

Semantic Parsing

Thomason, 2015

This proposal

Language Grounding

Thomason, 2016

Thomason, in progress
Outline

● Background

● Completed work

● Proposed Work
  ○ Synset Induction for Multi-modal Grounded Predicates
  ○ Grounding Semantic Parses Against Knowledge and Perception
  ○ Long-term Proposals

● Conclusion
Synset Induction for Multi-modal Grounded Predicates

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

- Perception information to identify referent object
  - Now we have methodology to identify senses of “light”
  - Need to integrate with “I, Spy” multi-modal perception
Synset Induction for Grounded Predicates

- In “I, Spy”, users used polysemous words like “light”
- Synset induction could combine the color-sense of “light” with “pale”, a rarer descriptor
  - mug1_cup2
- Expect synset-level classifiers to have cleaner positive examples (single-sense) and more of them (from multiple words)
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
Outline

● Background

● Completed work

● Proposed Work
  ○ Synset Induction for Multi-modal Grounded Predicates
  ○ Grounding Semantic Parses Against Knowledge and Perception
  ○ Long-term Proposals

● Conclusion
Grounding Semantic Parses against Knowledge and Perception

- **Commands** that need to be actualized through robot action
- **World knowledge** about people and the surrounding office space
- **Perception information** to identify referent object
- An integrated system of completed works could achieve all goals
- Creates new challenges
- Affords new opportunities for continuous learning

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

- In vanilla semantic parsing, all predicates are known in a given ontology.
- People may use words to express new concepts after the “I, Spy”-style bootstrapping phase.
- “Take that tiny box to Bob”
- Does unseen word tiny refer to a novel concept or existing synset?
- Unseen adjectives and nouns start as novel single-sense synsets.
- Synset induction can later collapse these to their synonyms (here, small).
- Other words, like pointy, may refer to formerly unseen concepts.
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”

0.6 light₁ mug₁
0.4 light₂ mug₁

re-ranking

0.6 * 0.3 * 0.8 = 0.144 light₁ mug₁
0.4 * 0.7 * 0.8 = 0.224 light₂ mug₁
```
Perception Training Data from Dialog

- “Bring me the light mug”
- Human can confirm correct object was delivered
- Then delivered object is positive example for $light_2$ and $mug_1$
Outline

● Background
● Completed work
● Proposed Work
  ○ Synset Induction for Multi-modal Grounded Predicates
  ○ Grounding Semantic Parses Against Knowledge and Perception
  ○ Long-term Proposals
● Conclusion
Intelligent Exploration of Novel Objects

● “get the pink marker”
● Don’t need to lift, drop, etc. a new object to determine whether it’s “pink”
● Can consult sensorimotor context classifiers for “pink” to determine which behaviors are most informative (e.g. look)
● Still need to lift objects to determine “heavy”
Positive-unlabeled Learning for Perception

- SVMs currently power sensorimotor context classifiers
- Require positive and negative object examples to make decisions
- Could swap these out for positive-unlabeled learning methods
- Only positive examples needed, so data could come from dialog alone
- Confirm referent object with human to get positive examples for predicates involved
Leveraging Accommodation

● Want humans and robots to communicate effectively
● Can try to modify human utterances in a natural way in addition to better understanding them
● Accommodation is a natural phenomenon
  ○ Lexical and syntactic agreement; pitch and loudness convergence
● Have dialog generate utterances it would understand well itself
● Tacitly encourage user to speak in ways the NLU better understands
Outline

- Background
- Completed work
- Proposed Work
- Conclusion
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
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 ✓
  - Even with polysemy
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 ✔
  - Even with polysemy ✔
Go to Alice’s office and get the light mug for the chair.
I will go to Room 1, pick up a light mug object, and deliver it to Bob.
Continuously Improving Natural Language Understanding for Robotic Systems through Semantic Parsing, Dialog, and Multi-modal Perception

Jesse Thomason
Doctoral Dissertation Proposal