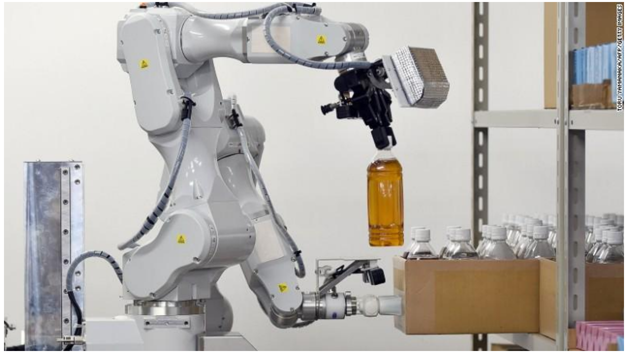
Improved Models and Queries for Grounded Human-Robot Dialog

Aishwarya Padmakumar

Doctoral Dissertation Proposal

Natural Language Interaction with Robots



Understanding Commands

Bring the blue mug
from Alice’s office

Sources of Imperfect Understanding

- Language is inherently ambiguous
  - Mug:  vs  vs 
- Imperfect models
  - Fail to detect the mug
- Missing domain specific knowledge
  - Alice’s office is missing in the directory

Dialog - Clarification

Bring the blue mug from Alice's office

Where should I bring a blue mug from?

Alice Ashcraft's office

I should bring a blue mug from 3502?

Yes

Dialog - Improve Models

Bring the blue mug from Alice's office

Where should I bring a blue mug from?

Alice Ashcraft's office

I should bring a blue mug from 3502?

Yes

Alice's office

≍

Alice Ashcraft's office

≍

3502

Dialog - Acquiring Labels

Bring the blue mug from Alice's office

Would you use the word "blue" to refer to this object?

Yes

This Proposal

Improving grounded human-robot dialog by

- Learning dialog policies from interactions
- Improved queries to be used in dialogs
- Improved models for perceptual grounding

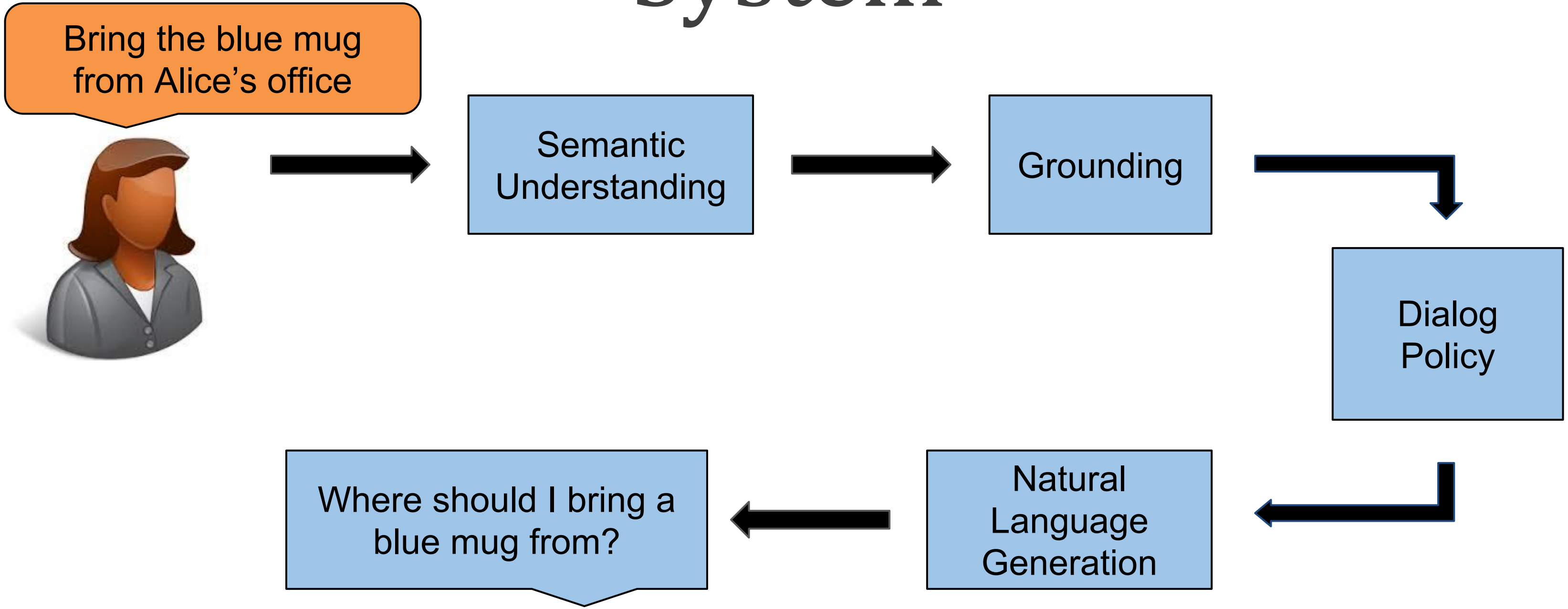
Outline

- Background
- Completed Work
  - Integrating Learning of Dialog Strategies and Semantic Parsing (Padmakumar et.al., 2017)
  - Opportunistic Active Learning for Grounding Natural Language Descriptions (Thomason et. al., 2017)
  - Learning a Policy for Opportunistic Active Learning (Padmakumar et. al., 2018)
- Proposed Work
- Conclusion

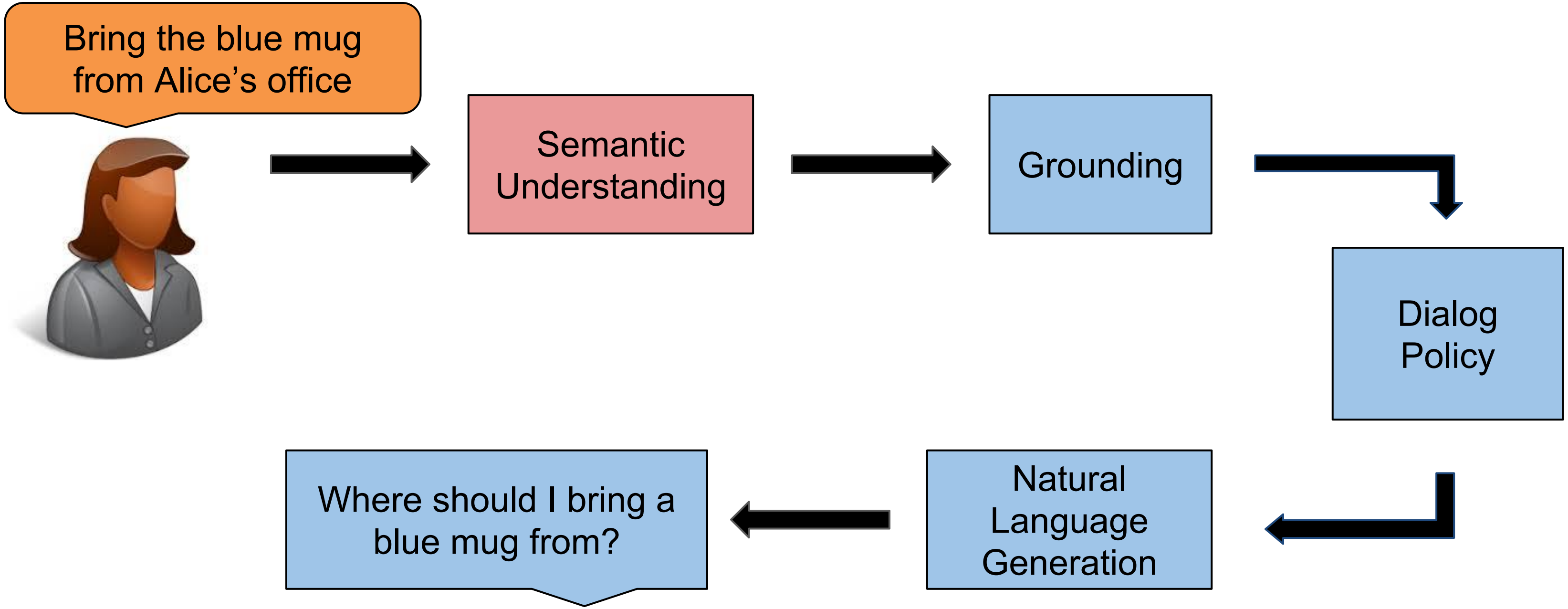
Outline

- Background
- Completed Work
  - Integrating Learning of Dialog Strategies and Semantic Parsing (Padmakumar et.al., 2017)
  - Opportunistic Active Learning for Grounding Natural Language Descriptions (Thomason et. al., 2017)
  - Learning a Policy for Opportunistic Active Learning (Padmakumar et. al., 2018)
- Proposed Work
- Conclusion

Background: Parts of a Dialog System



Background: Semantic Understanding



Background: Semantic Understanding

Convert natural language into a machine understandable representation

Background: Semantic Understanding

Convert natural language into a machine understandable representation

Bring the blue mug from Alice's office

$$\texttt{bring}(\lambda x.(\texttt{blue}(x) \wedge \texttt{mug}(x)), \\ \texttt{the}(\lambda y.(\texttt{office}(y) \wedge \texttt{owns}(\texttt{alice}, y))))$$

Semantic parsing -

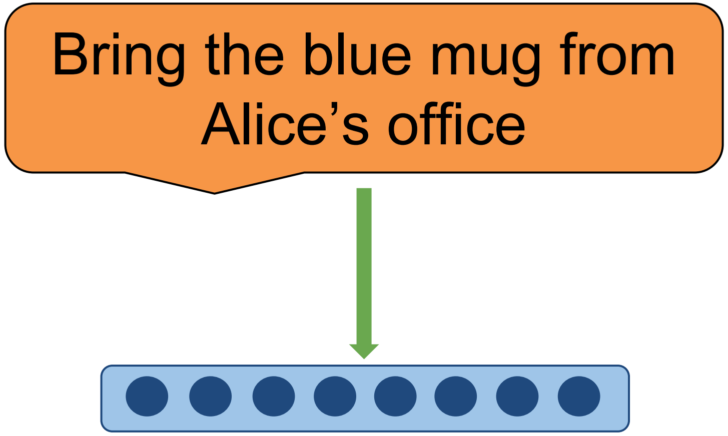
- Converts language to a structured meaning representation
- Compositionality - meaning of "*blue mug*" from meaning of "*blue*" and meaning of "*mug*"

Background: Semantic Understanding

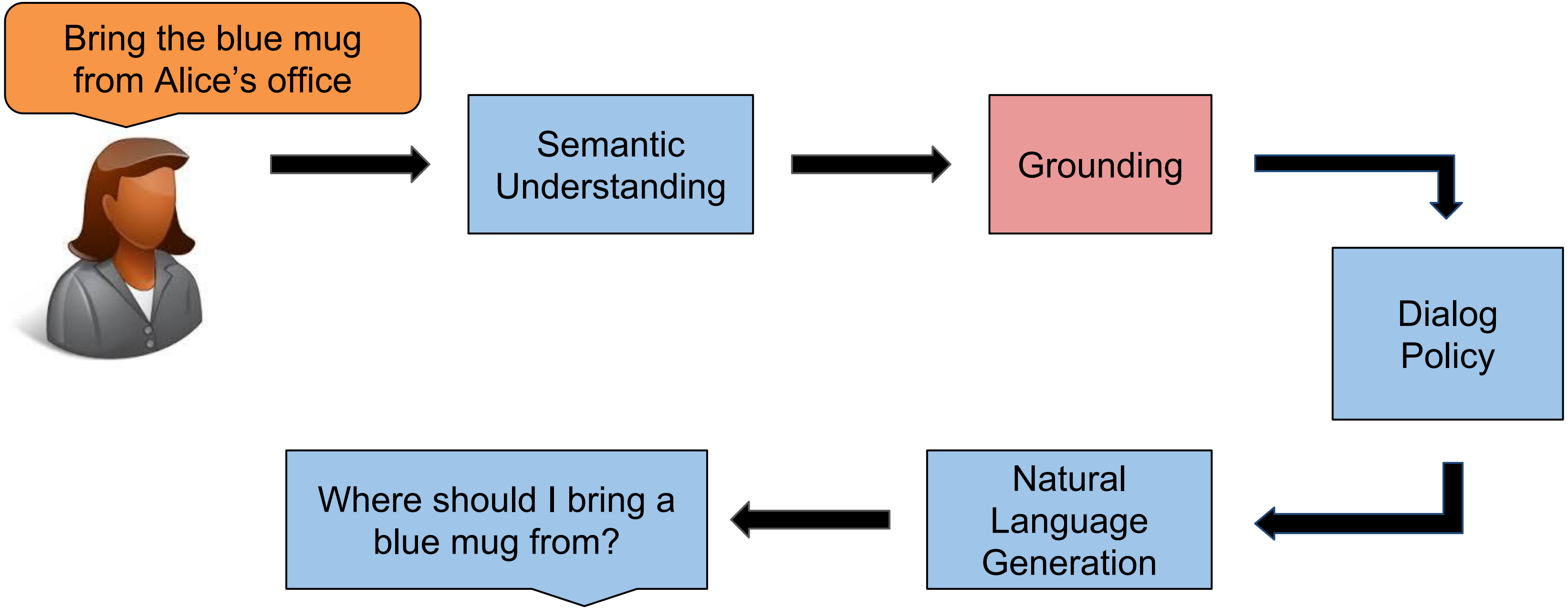
Convert natural language into a machine understandable representation

Vector Space Representations -

- Converts words/sentences to vectors that represent meaning.
- Typically non compositional.
- Less initial handcrafting
- More training data



Background: Grounding



Background: Grounding

Map meaning representations to real world entities

Background: Grounding

Map meaning representations to real world entities

$$\texttt{the}(\lambda y.(\texttt{office}(y) \wedge \texttt{owns}(\texttt{alice}, y)))$$

Knowledge Base Grounding

Person	Office
alice	3502
bob	3324

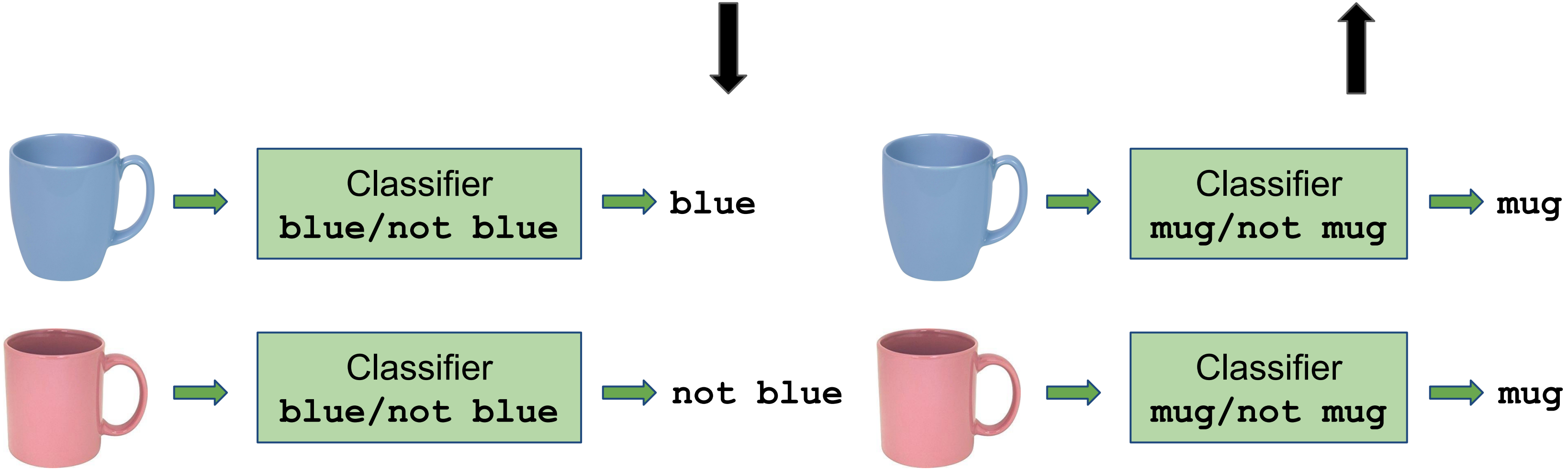
`3502`

Background: Grounding

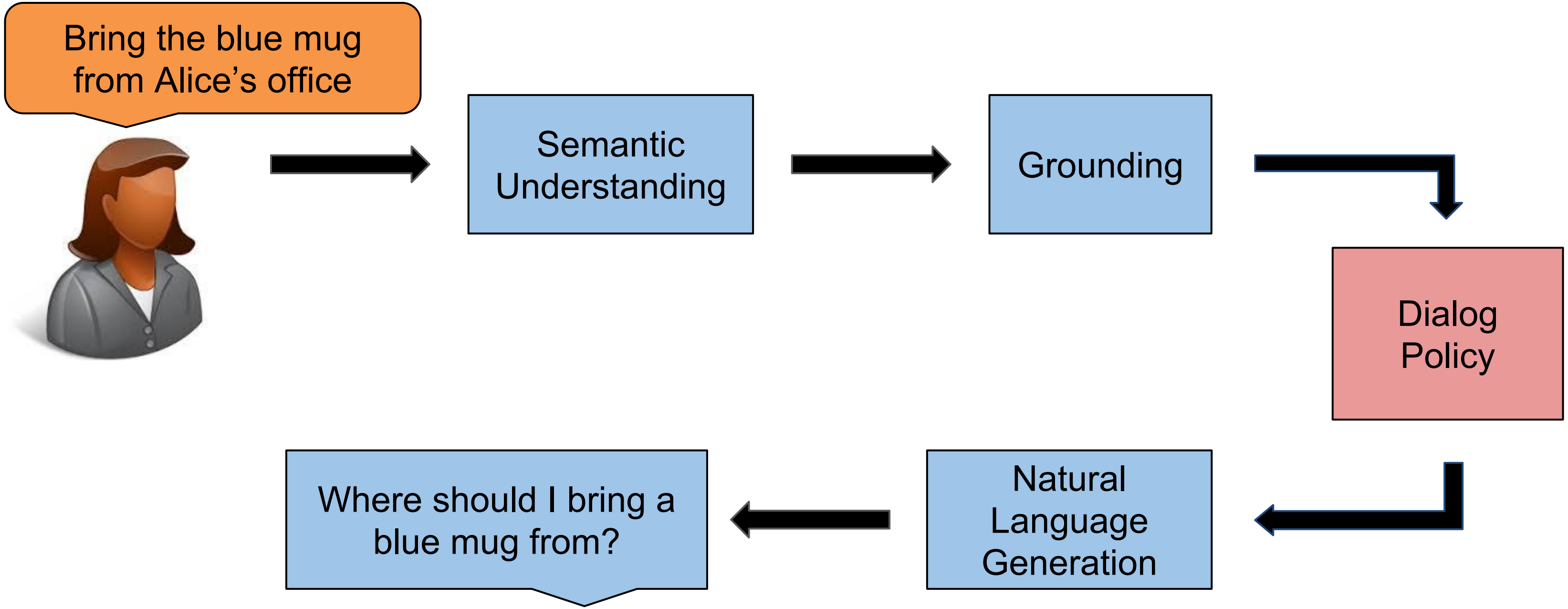
Map meaning representations to real world entities

Perceptual Grounding

$$\lambda x.(\mathtt{blue}(x) \wedge \mathtt{mug}(x))$$



Background: Dialog Policy

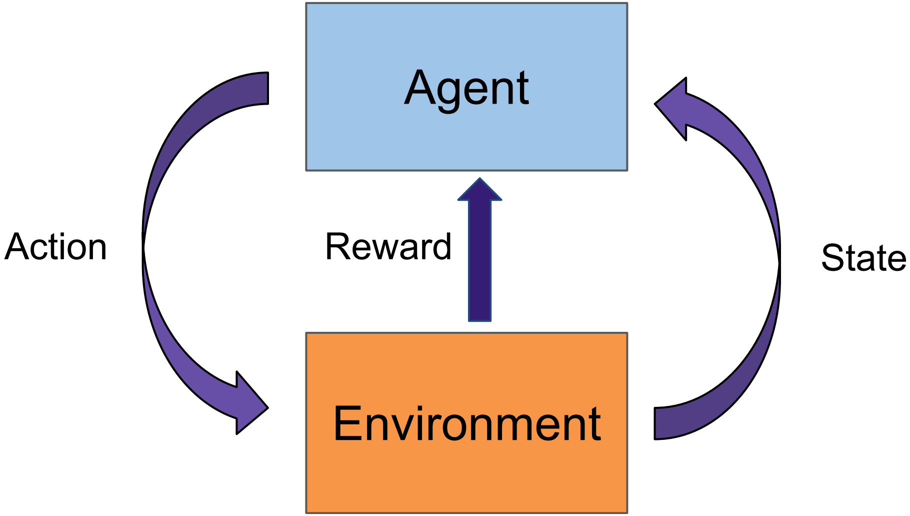


Background: Dialog Policy

- Decides each response type - clarification, label queries, task completion
- Dialog state - Information from the dialog so far
- Dialog policy - Mapping from dialog states to dialog actions (response types/ responses)
- Learned using Reinforcement Learning

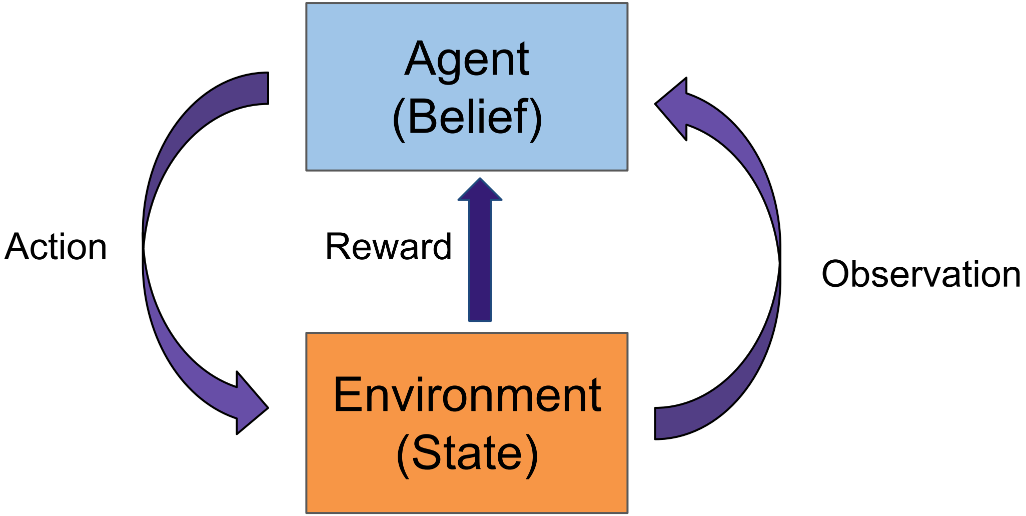
Background: Reinforcement Learning

Markov Decision Process (MDP)

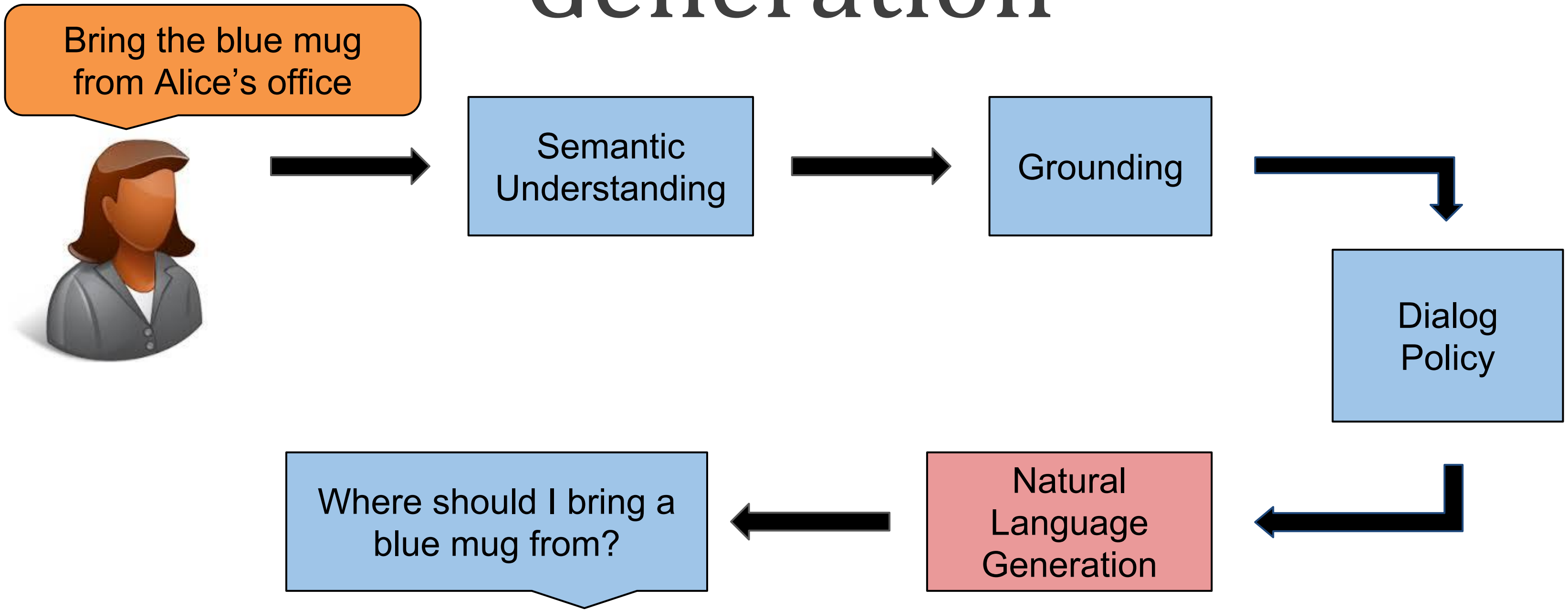


Background: Reinforcement Learning

Partially Observable Markov Decision Process (POMDP)



Background: Natural Language Generation



Background: Natural Language Generation

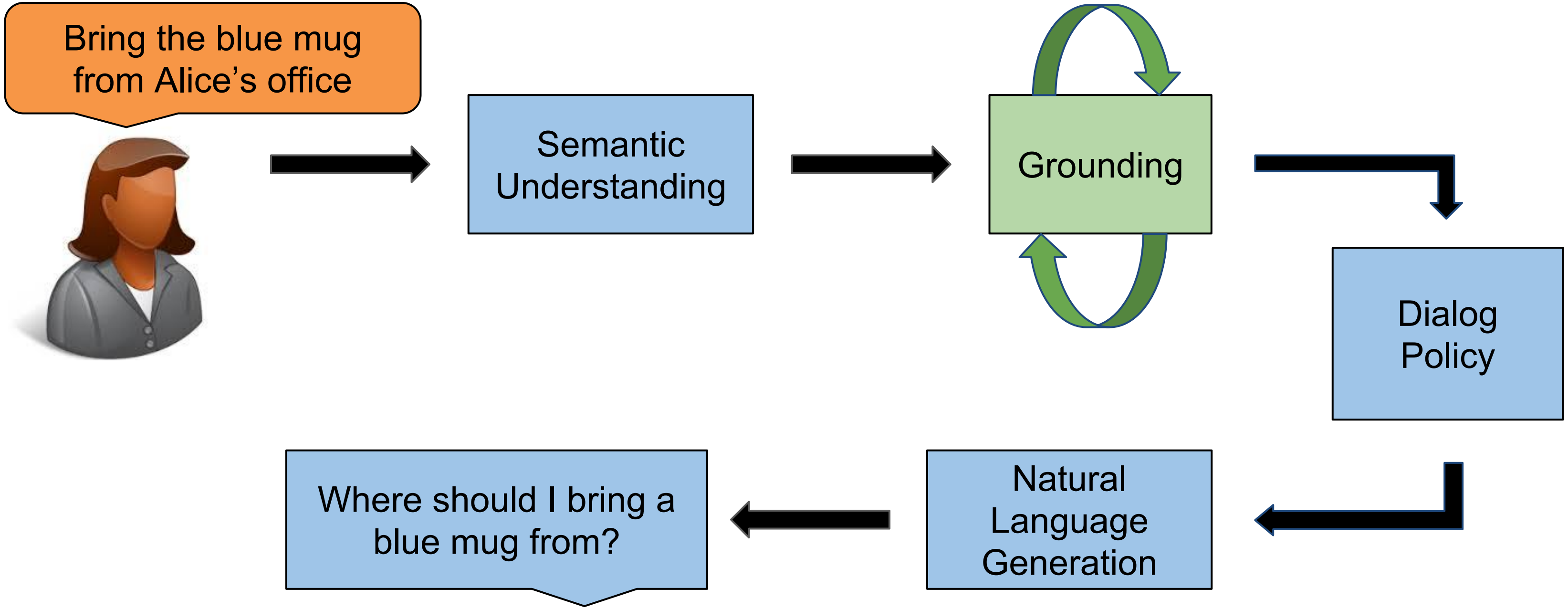
Converting an action to a natural language response

```
ask_param(
    action=bring,
    patient=λx.(blue(x) ∧ mug(x))
    src=?
)
```

→

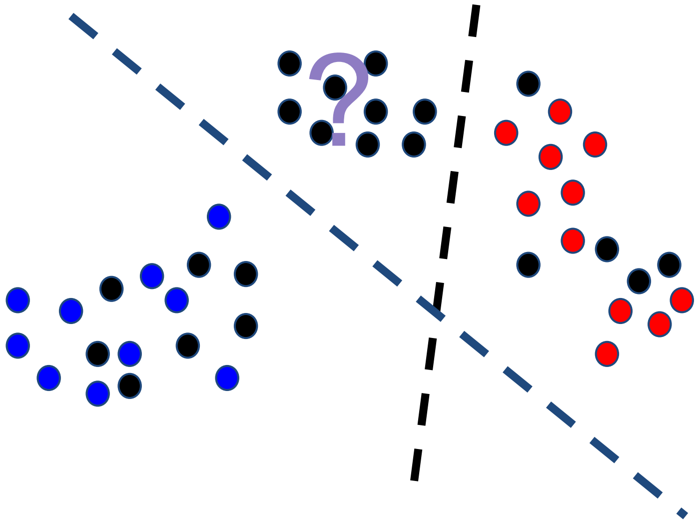
Where should I bring a blue mug from?

Background: Active Learning



Background: Active Learning

Query for labels most likely to improve the model.



Background: Active Learning

Bring the blue mug from Alice's office

Would you use the word "blue" to refer to this object?

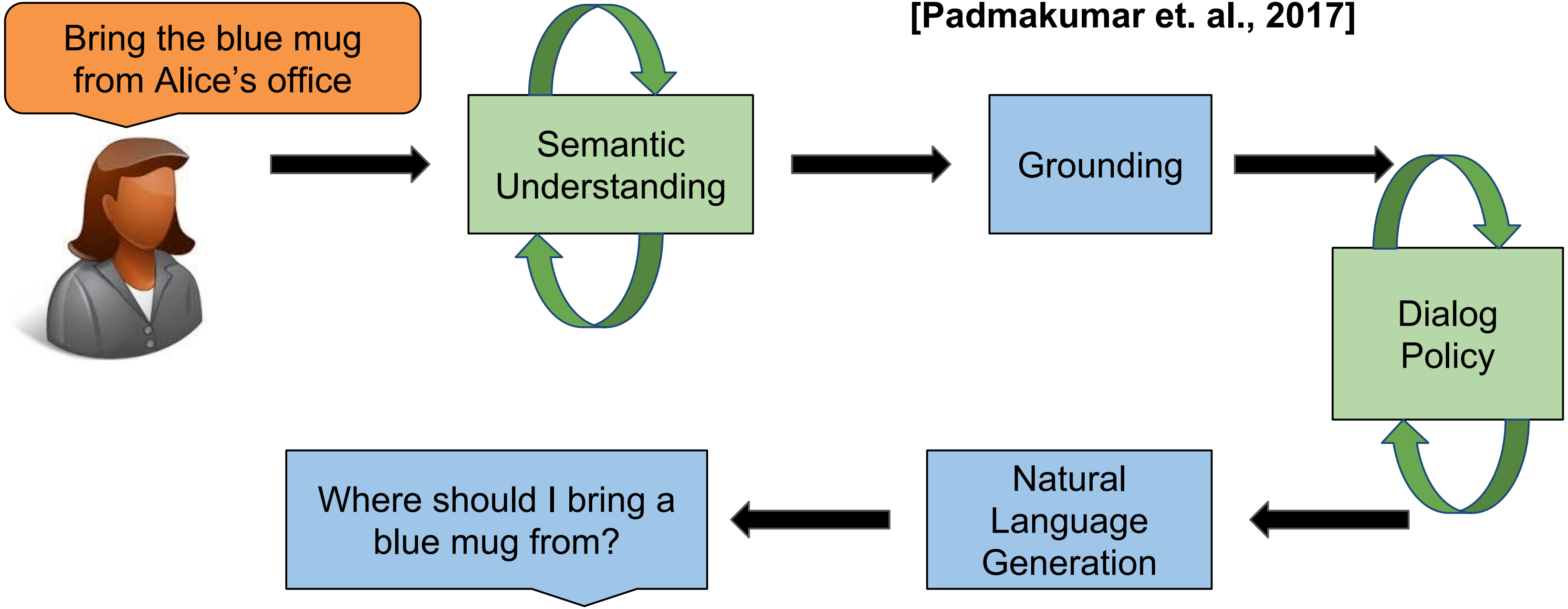
Yes

Outline

- Background
- Completed Work
  - Integrating Learning of Dialog Strategies and Semantic Parsing (Padmakumar et.al., 2017)
  - Opportunistic Active Learning for Grounding Natural Language Descriptions (Thomason et. al., 2017)
  - Learning a Policy for Opportunistic Active Learning (Padmakumar et. al., 2018)
- Proposed Work
- Conclusion

Integrating Learning of Dialog Strategies and Semantic Parsing

[Padmakumar et. al., 2017]



Prior work: Improving Semantic Parsers from Clarification Dialogs

[Thomason et. al., 2015]

Bring the blue mug from Alice's office

Where should I bring a blue mug from?

Alice Ashcraft's office

I should bring a blue mug from 3502?

Yes

Alice's office

$\asymp$

Alice Ashcraft's office

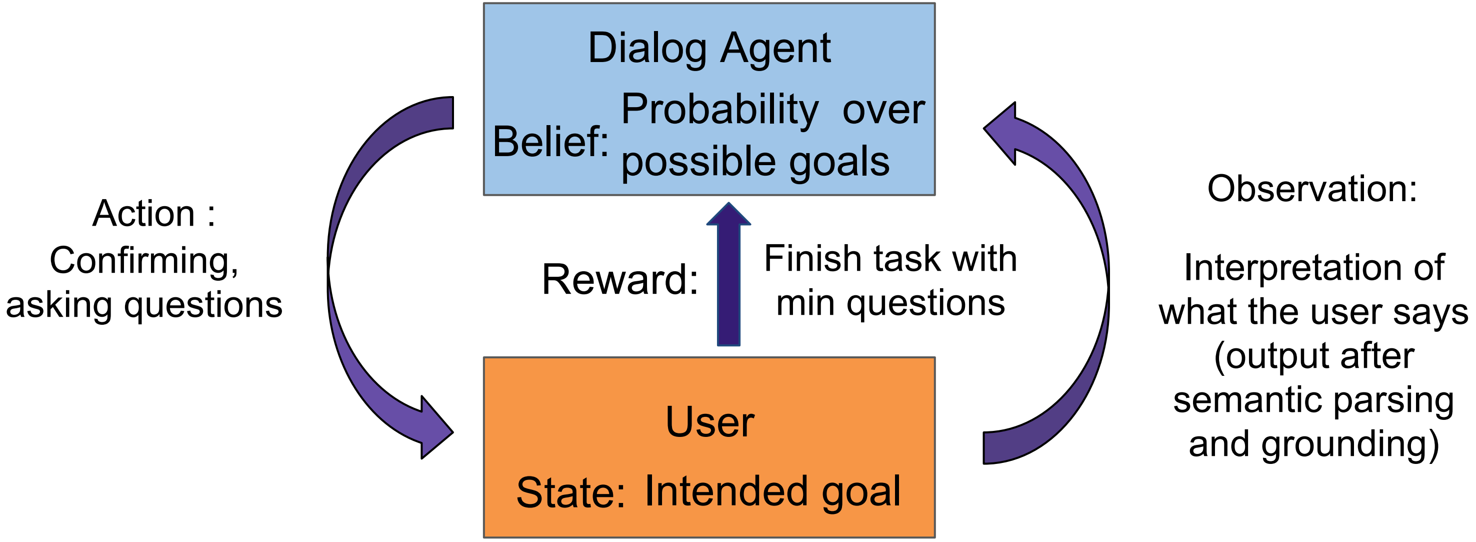
$\asymp$

3502

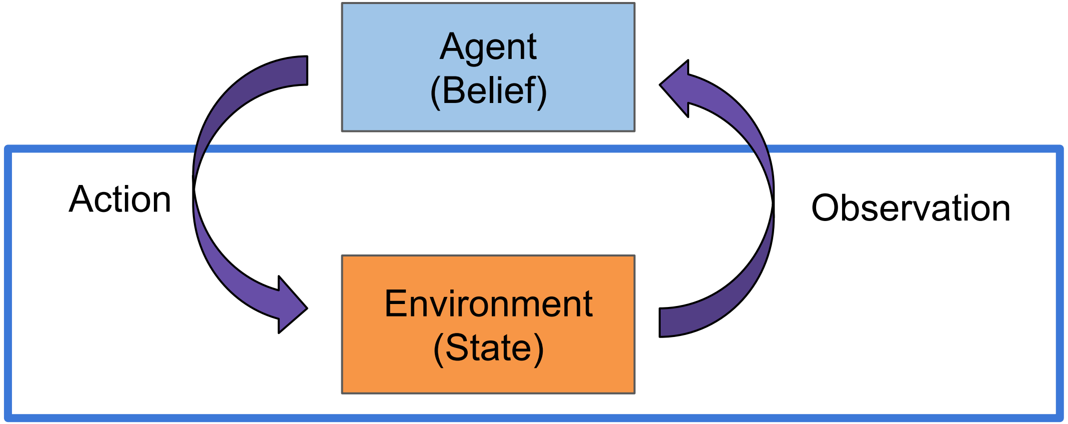
Prior Work: Dialog Policy Learning

Modelling dialog system as a Partially Observable Markov Decision Process (POMDP)

[Young et. al., 2013]

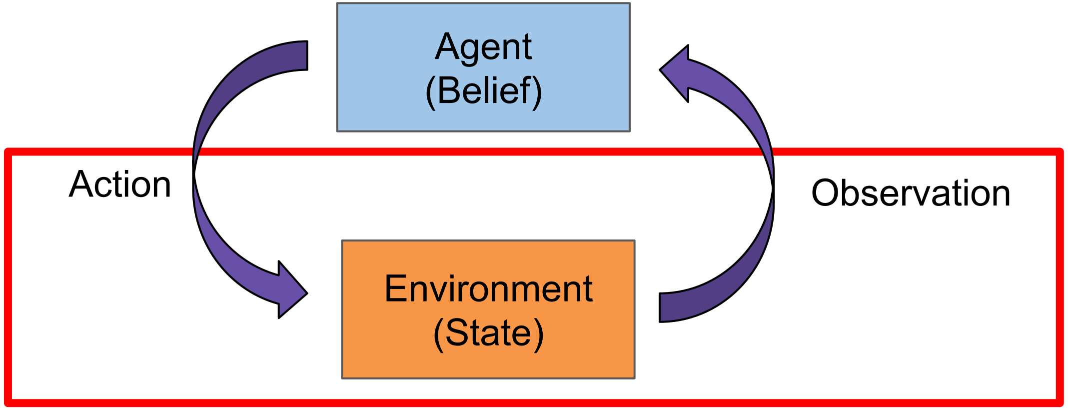


Why is joint learning challenging?



Assumption:

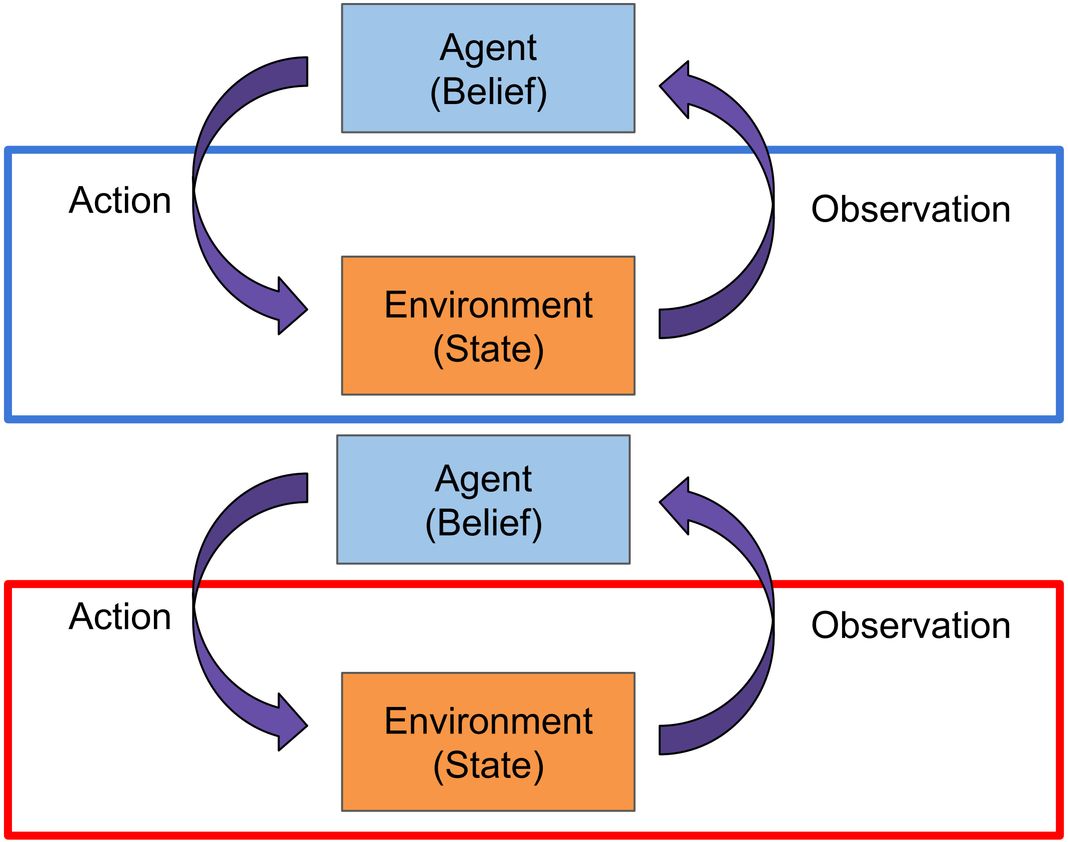
Constant probability distribution



Our system:

Variable probability distribution

Why is joint learning challenging?



Assumption:

Constant probability distribution

Non-stationary Environment

Choosing a Policy Learning Algorithm

- Robust to non-stationary environment - to allow simultaneous learning of a semantic parser
- Learns how the mapping from states and actions to observations varies with time
- Low Sample Complexity - Learn a good policy from a small number of dialogs
- Kalman Temporal Differences (KTD) Q Learning (Geist and Pietquin, 2010)

Experiments - Mechanical Turk

TASK TO COMPLETE

Dave Daniel wants the item in slot 5.

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

Final Task

DIRECTORY

People:

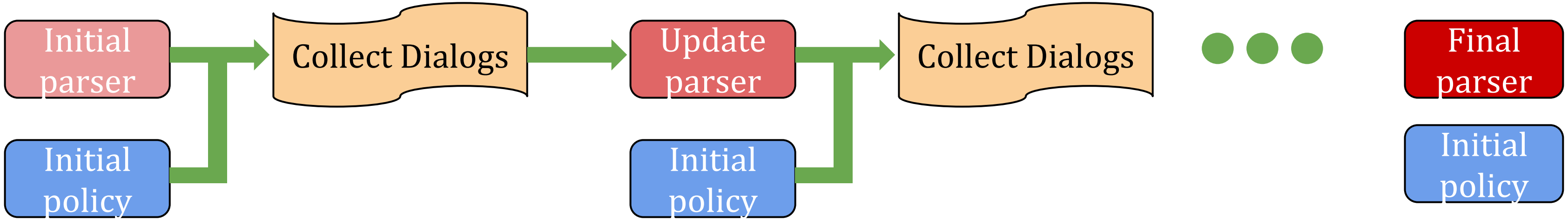
Alice Ashcraft; Secretary
Francis (“Frannie”) Foster
Robert (“Bob”) Brown
Carol Clark, PhD
Dave Daniel, PhD
George Green; Intern
Evelyn (“Eve”) Eckhart
Mallory Morgan; Director
Peggy Parker, PhD
Walter Ward; Supervisor

Items available to robot:

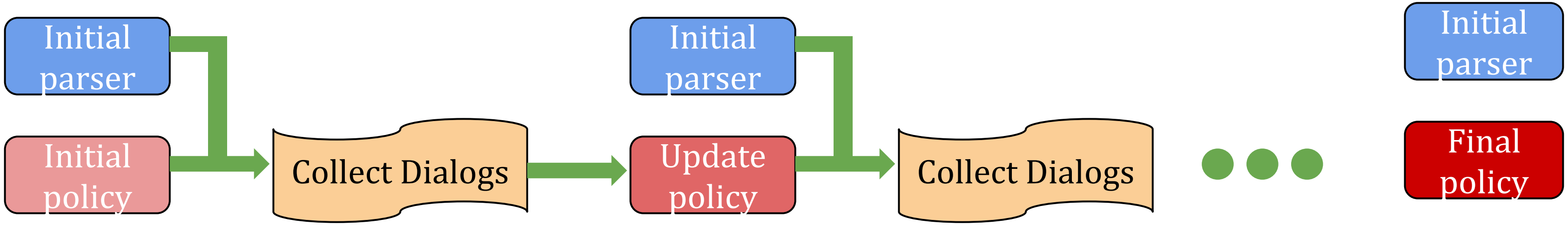
Image Source: Thomason et al., 2015

Experimental Conditions

Parser Learning

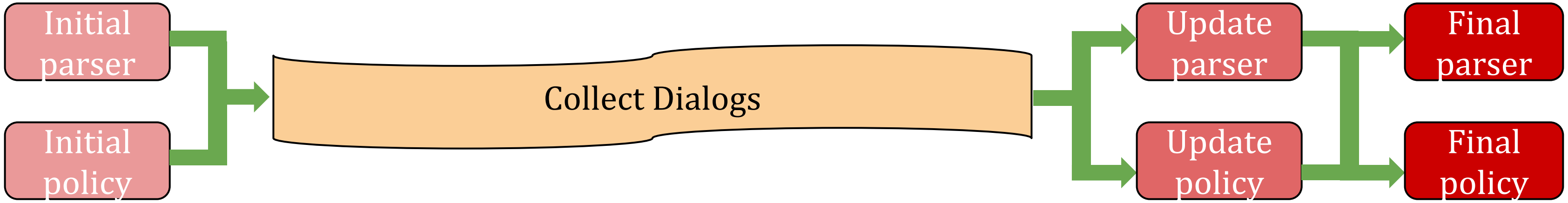


Dialog Learning

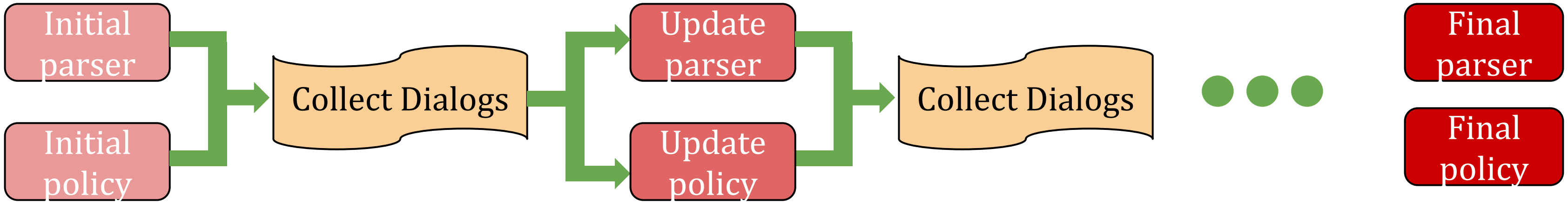


Experimental Conditions

Parser and Dialog Learning - Full (Naive)

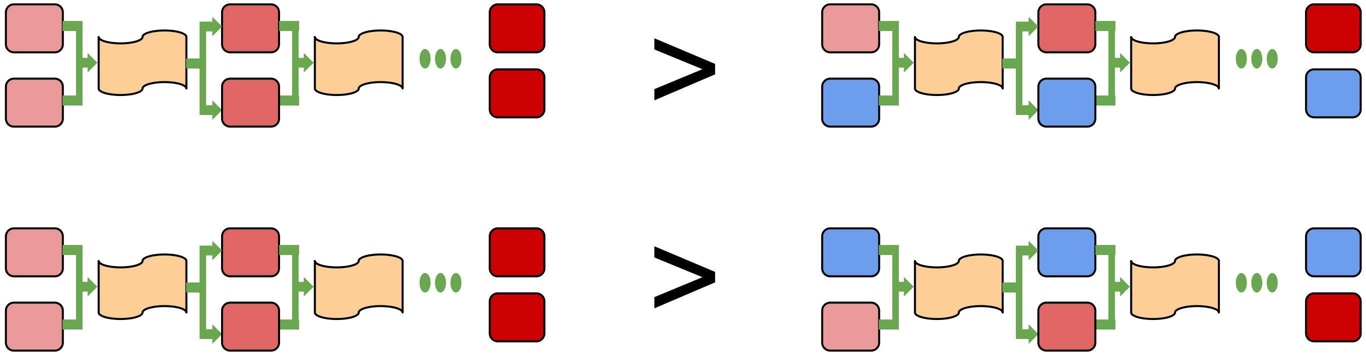


Parser and Dialog Learning - Batchwise (Ours)



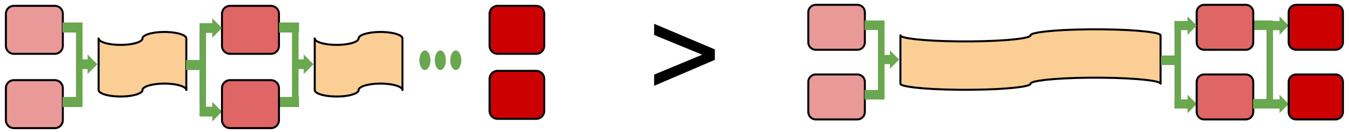
Hypotheses

1. Combined parser and dialog learning is more useful than either alone.

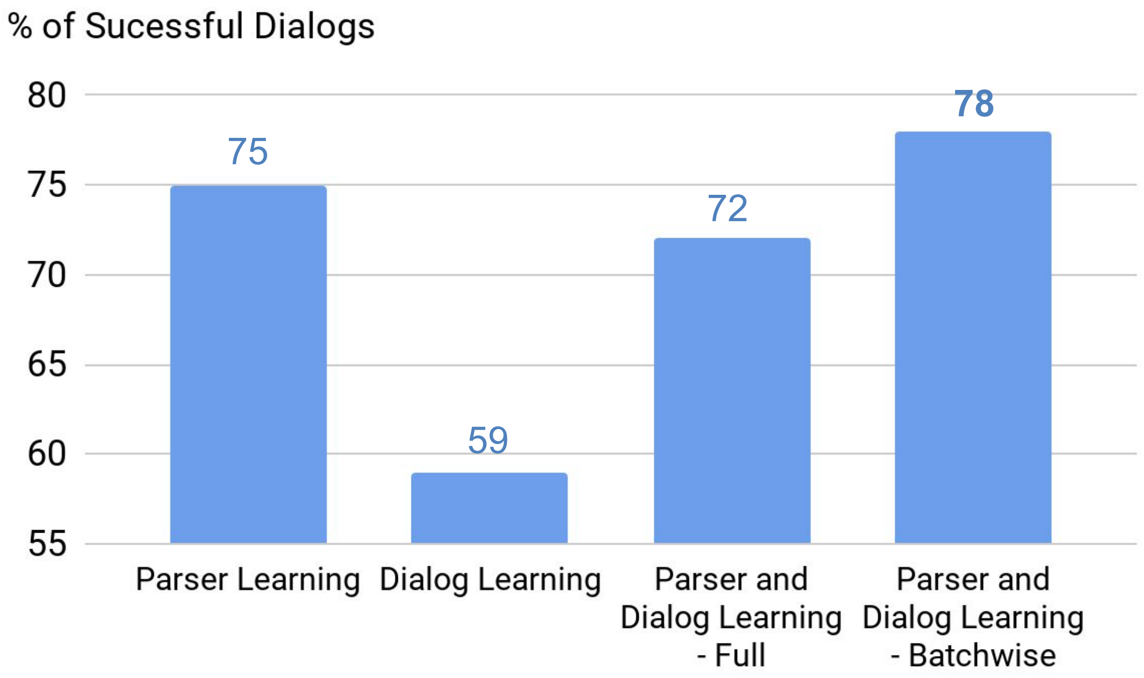


Hypotheses

2. Changes in the parser need to be seen by the dialog management module.

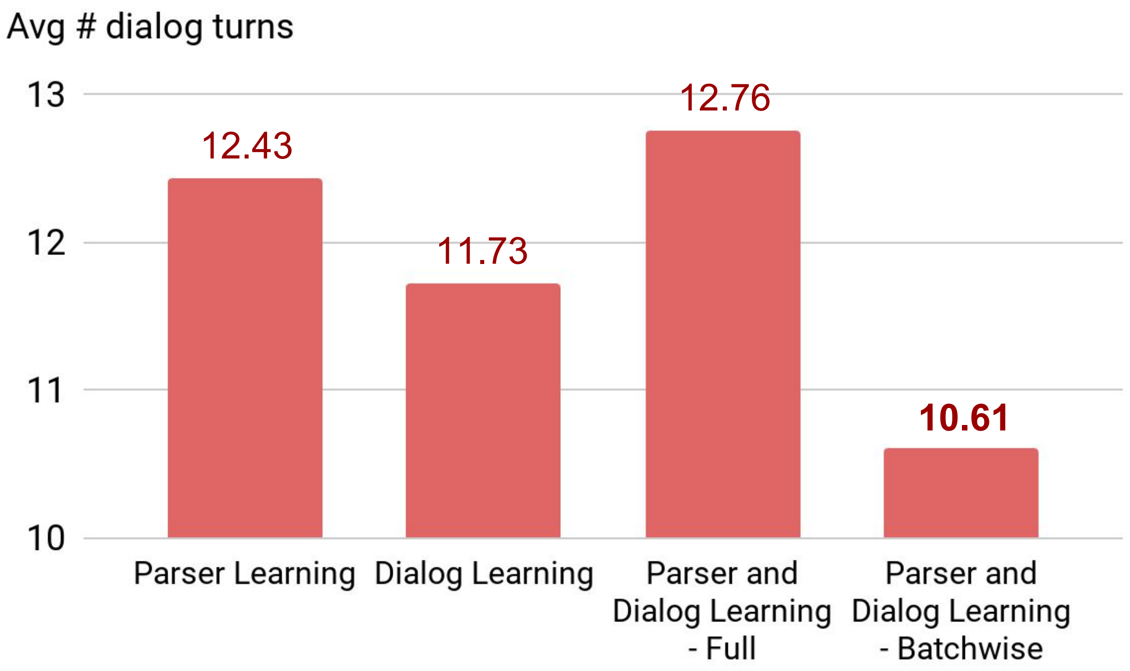


Results - Dialog Success



- Higher is better
- Parser learning is mostly responsible for improvement in dialog success rate
- Best system: parser and dialog learning - batchwise

Results - Dialog Length



- Lower is better
- Dialog learning is mostly responsible for lowering dialog length
- Best system: parser and dialog learning - batchwise

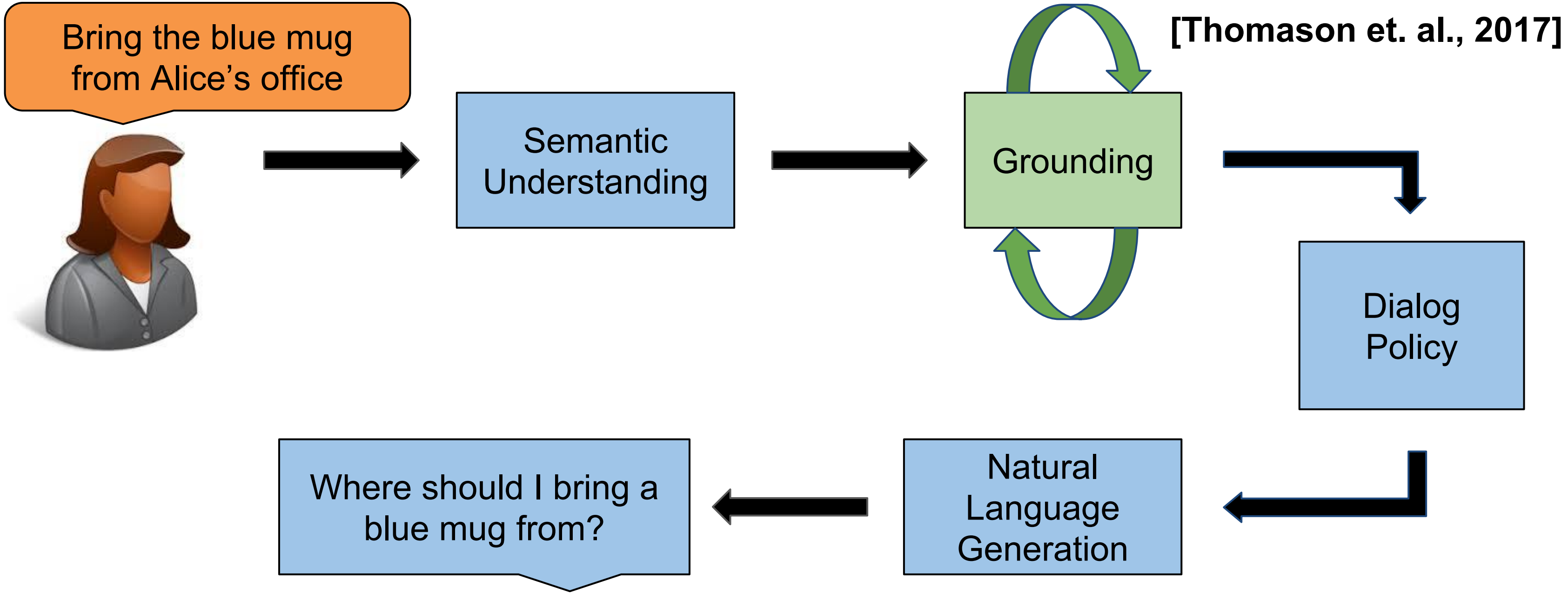
Conclusion

- Jointly learning a parser and dialog policy is more effective than learning either alone - qualitative and quantitative.
- Changes in other components need to be propagated to the policy.

Outline

- Background
- Completed Work
  - Integrating Learning of Dialog Strategies and Semantic Parsing (Padmakumar et.al., 2017)
  - Opportunistic Active Learning for Grounding Natural Language Descriptions (Thomason et. al., 2017)
  - Learning a Policy for Opportunistic Active Learning (Padmakumar et. al., 2018)
- Proposed Work
- Conclusion

Opportunistic Active Learning for Grounding Natural Language Descriptions



Opportunistic Active Learning

- Asking locally convenient questions during an interactive task.
- Questions may not be useful for the current interaction but expected to help future tasks.

Opportunistic Active Learning

Bring the blue mug from Alice’s office

Would you use the word “blue” to refer to this object?

Yes

Opportunistic Active Learning

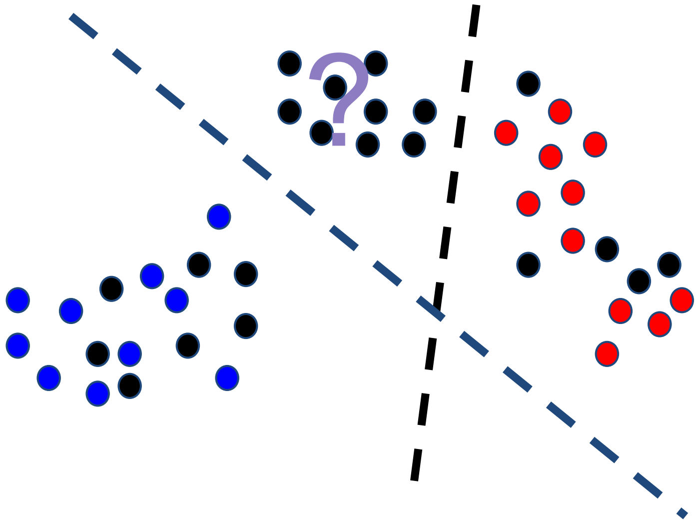
Bring the **blue mug** from Alice's office

Would you use the word "**tall**" to refer to this object?

Yes

Opportunistic Active Learning

Still query for labels most likely to improve the model.



Opportunistic Active Learning

Why?

- Robot may have good models for on-topic concepts.
- No useful on-topic queries.
- Some off-topic concepts may be more important because they are used in more interactions.

Opportunistic Active Learning - Challenges

Some other object might be a better candidate for the question

Purple?

Opportunistic Active Learning - Challenges

The question interrupts another task and may be seen as unnatural

Bring the **blue mug** from Alice's office

Would you use the word "**tall**" to refer to this object?

Opportunistic Active Learning - Challenges

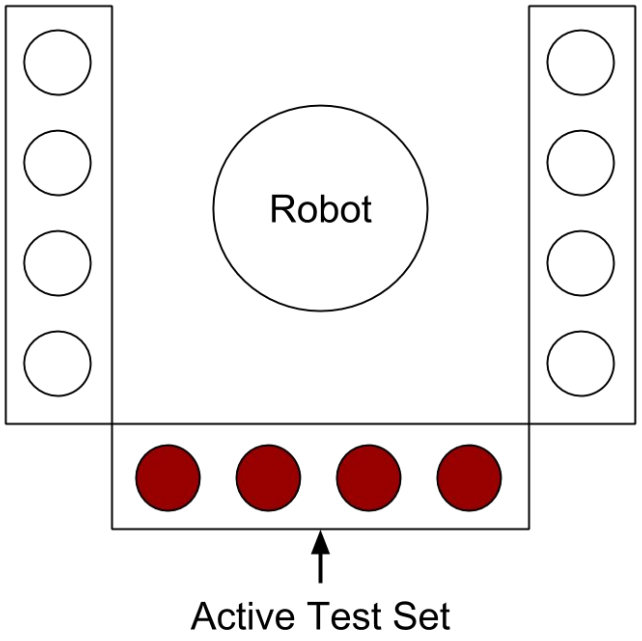
The information needs to be useful for a future task.

Red?

Object Retrieval Task

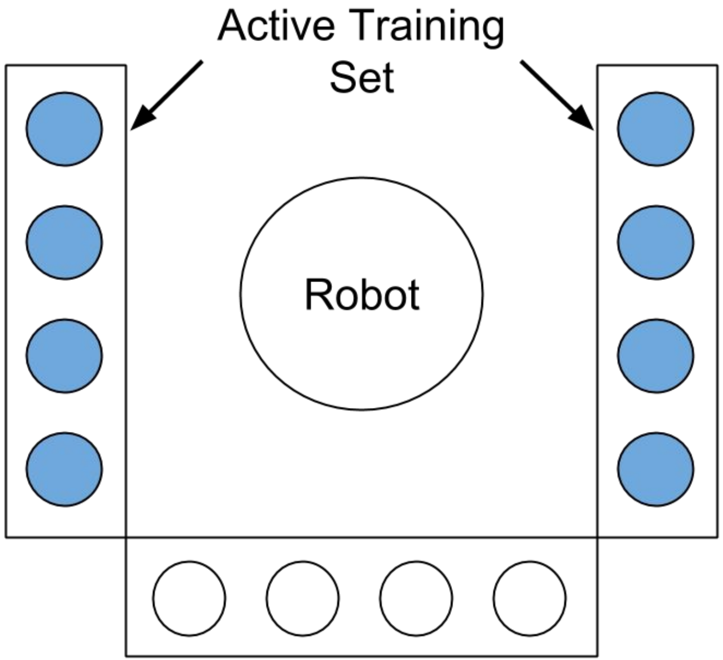
Object Retrieval Task

- User describes an object in the active test set
- Robot needs to identify which object is being described

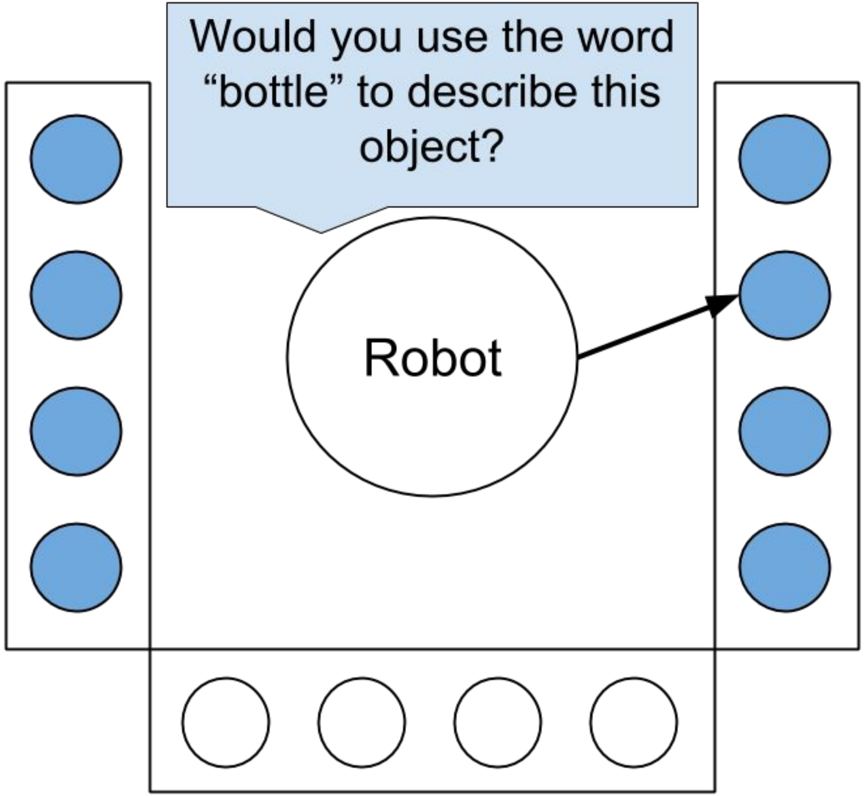


Object Retrieval Task

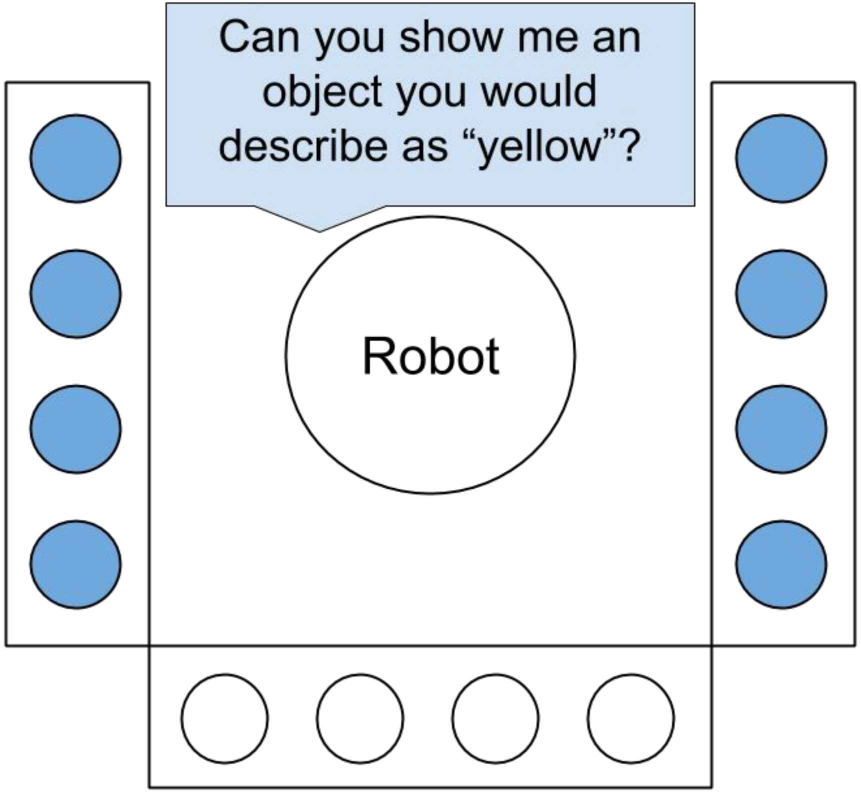
- Robot can ask questions about objects on the sides to learn object attributes



Two Types of Questions



Two Types of Questions



Experimental Conditions

This is a yellow bottle with water filled in it

- Baseline (on-topic) - the robot can only ask about “yellow”, “bottle”, “water”, “filled”
- Inquisitive (opportunistic) - the robot can ask about any concept it knows, possibly “red” or “heavy”

Results

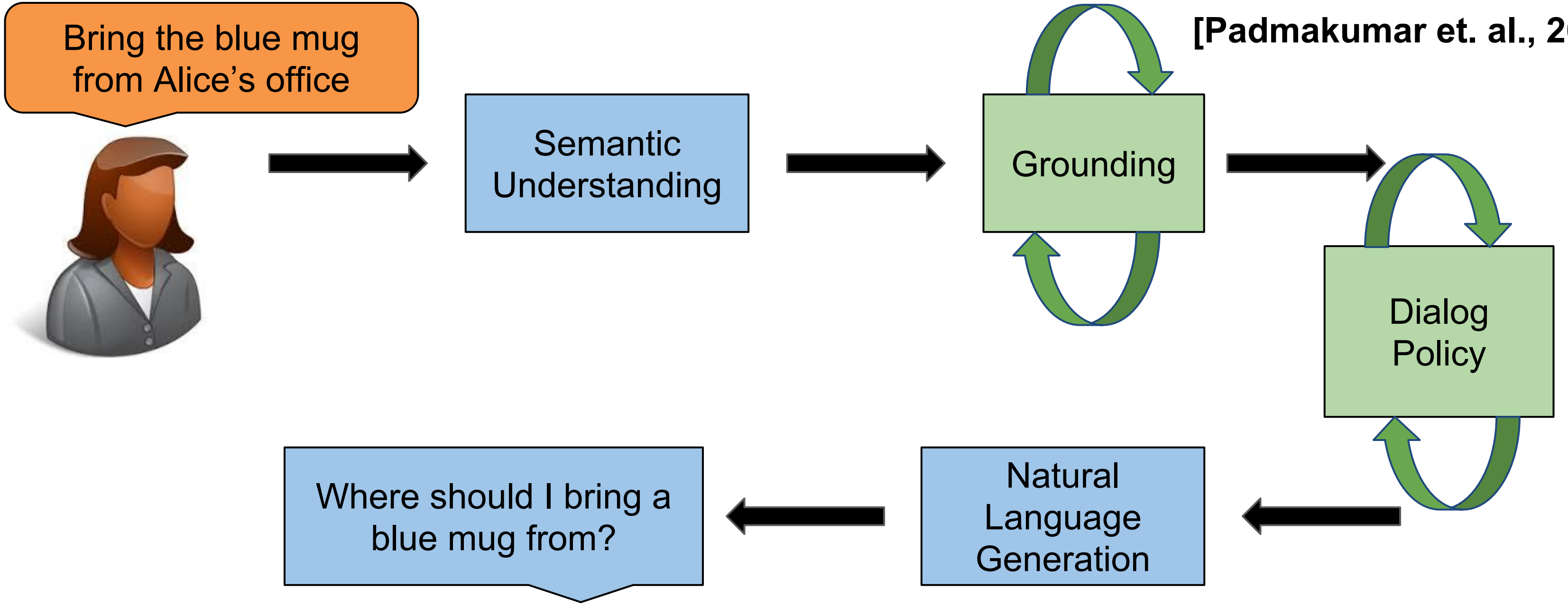
- Inquisitive robot performs better at understanding object descriptions.
- Users find the robot more comprehending, fun and usable in a real-world setting, when it is opportunistic.

Outline

- Background
- Completed Work
  - Integrating Learning of Dialog Strategies and Semantic Parsing (Padmakumar et.al., 2017)
  - Opportunistic Active Learning for Grounding Natural Language Descriptions (Thomason et. al., 2017)
  - Learning a Policy for Opportunistic Active Learning (Padmakumar et. al., 2018)
- Proposed Work
- Conclusion

Learning a Policy for Opportunistic Active Learning

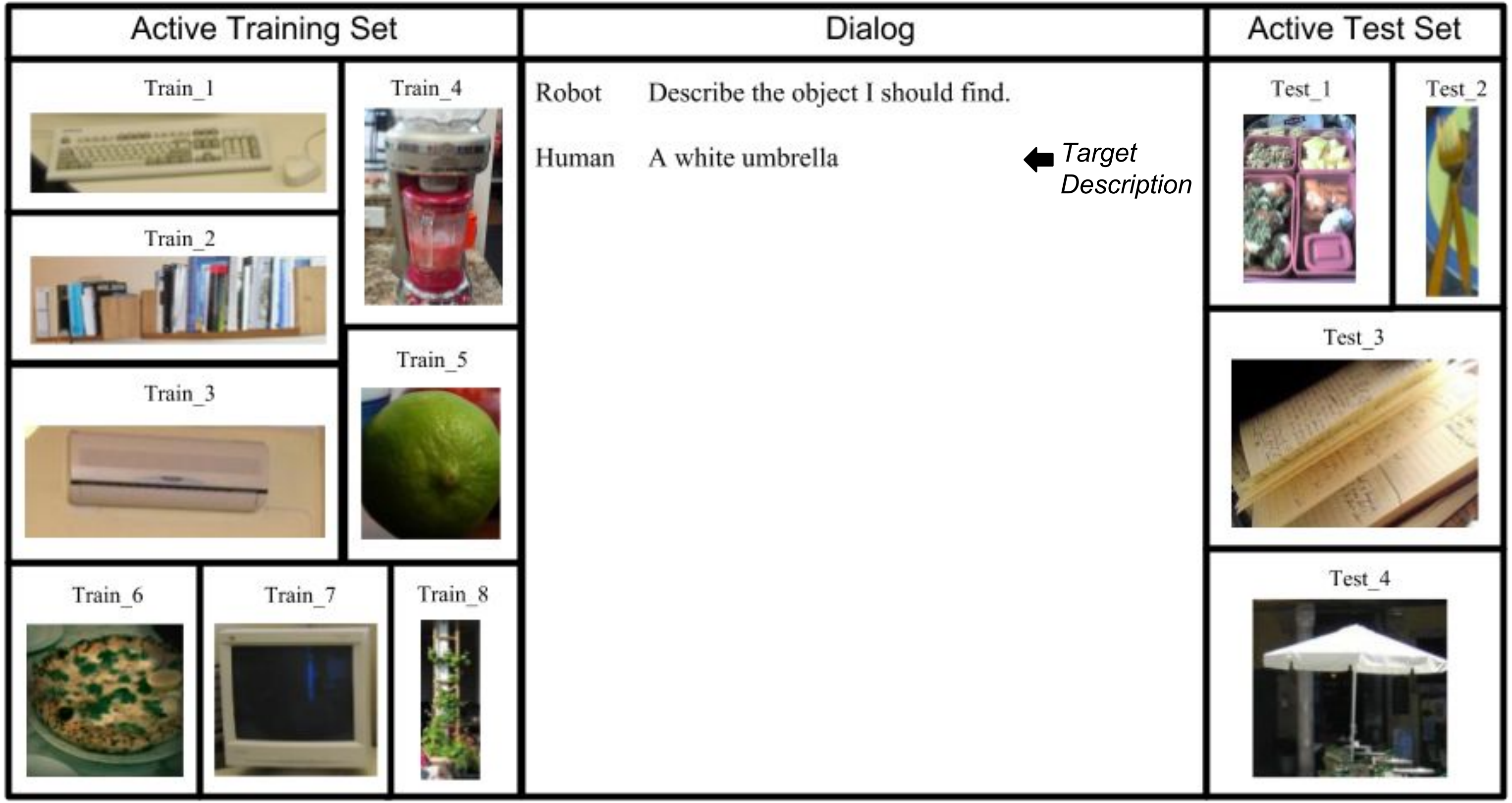
[Padmakumar et. al., 2018]



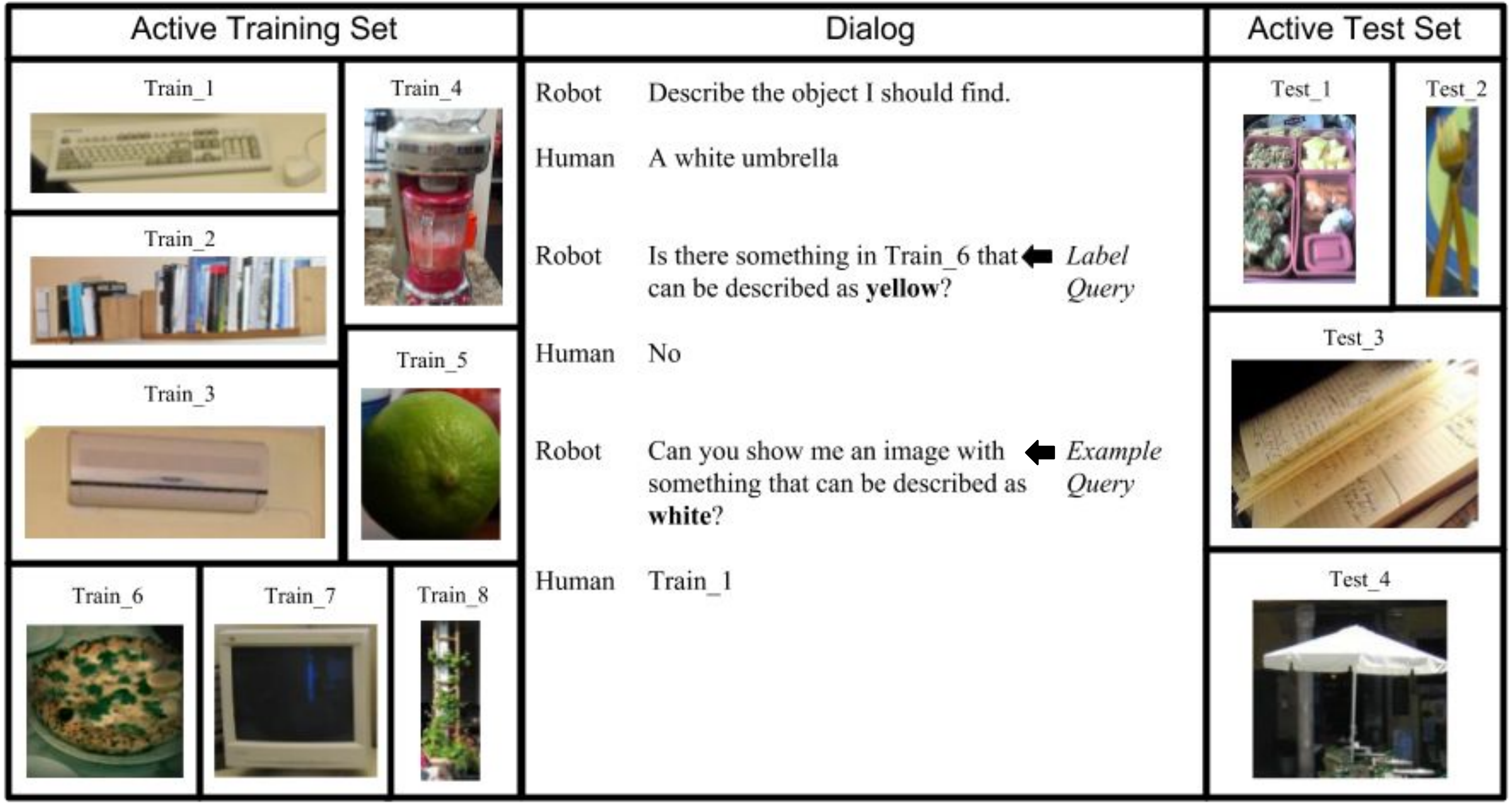
Learning a Policy for Opportunistic Active Learning

- Goal of this work - Learn a dialog policy that decides how many and which questions to ask to improve grounding models.
- To learn an effective policy, the agent needs to learn
  - To identify good queries in the opportunistic setting.
  - When a guess is likely to be successful.
  - To trade off between model improvement and task completion.

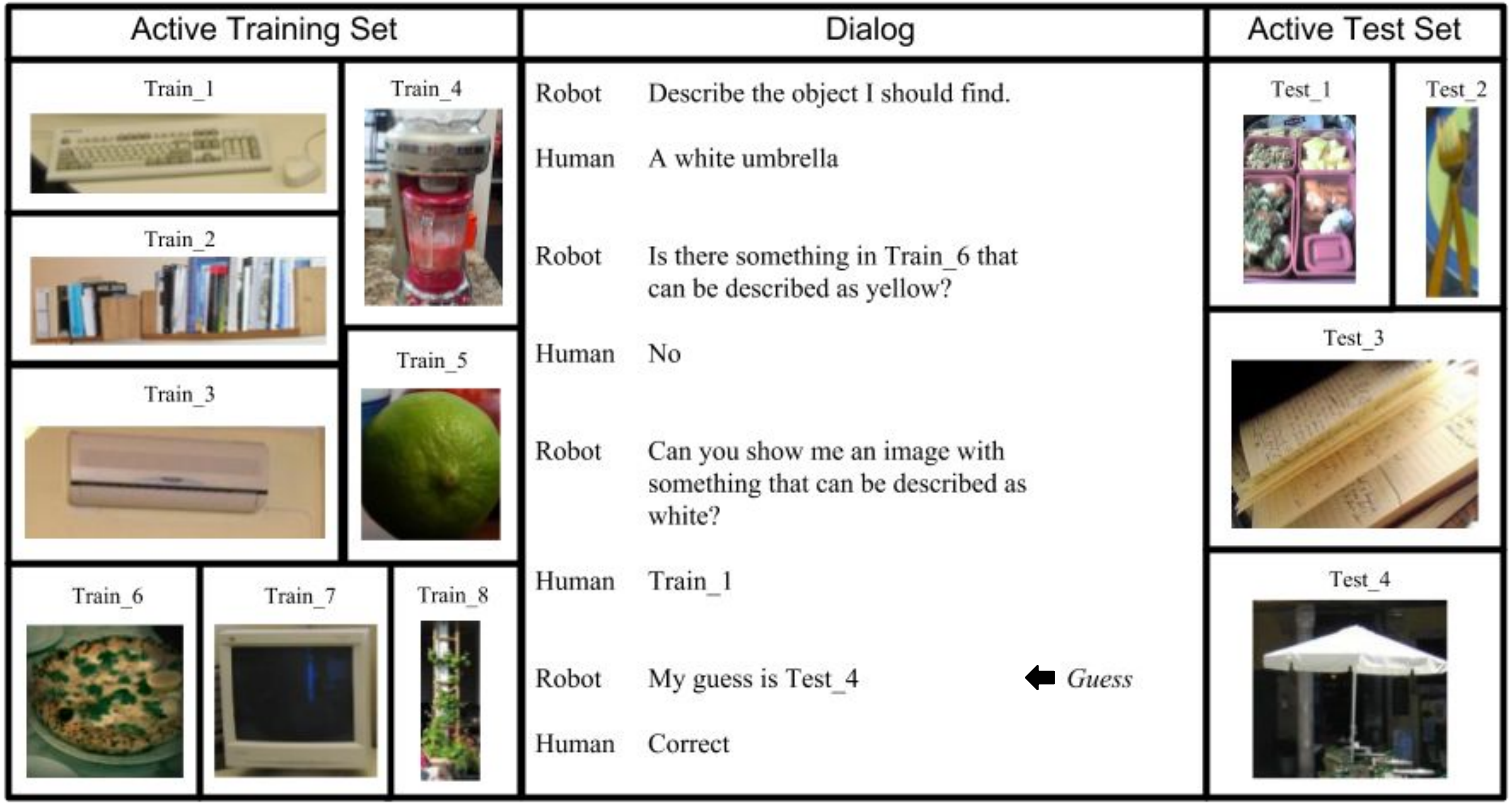
Task Setup


Active Training Set
Dialog
Active Test Set
Train_1
Train_2
Train_3
Train_4
Train_5
Train_6
Train_7
Train_8
Robot Describe the object I should find.
Human A white umbrella
Target Description
Test_1
Test_2
Test_3
Test_4

Task Setup


Active Training Set
Dialog
Active Test Set
Train_1
Train_2
Train_3
Train_4
Train_5
Train_6
Train_7
Train_8
Robot Describe the object I should find.
Human A white umbrella
Robot Is there something in Train_6 that can be described as **yellow**?
Label Query
Human No
Robot Can you show me an image with something that can be described as **white**?
Example Query
Human Train_1
Test_1
Test_2
Test_3
Test_4

Task Setup


Active Training Set
Dialog
Active Test Set
Train_1
Train_2
Train_3
Train_4
Train_5
Train_6
Train_7
Train_8
Robot Describe the object I should find.
Human A white umbrella
Robot Is there something in Train_6 that can be described as yellow?
Human No
Robot Can you show me an image with something that can be described as white?
Human Train_1
Robot My guess is Test_4
Guess
Human Correct
Test_1
Test_2
Test_3
Test_4

Grounding Model

A white umbrella → {white, umbrella}

Pretrained CNN

SVM → white/ not white

SVM → umbrella/ not umbrella

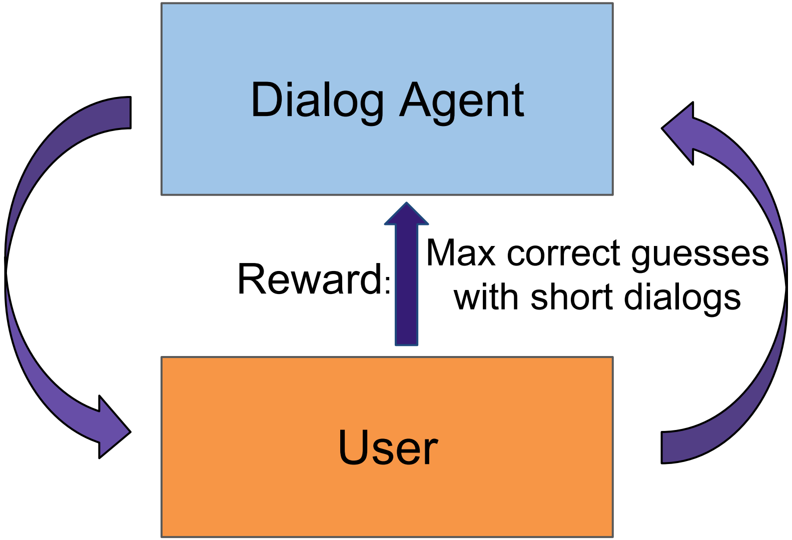
Active Learning

- Agent starts with no classifiers.
- Labeled examples are acquired through questions and used to train the classifiers.
- Agent needs to learn a policy to balance active learning with task completion.

MDP Model

Action:

- Label query
- Example Query
- Guess



State:

- Target description
- Train and test objects
- Agent’s perceptual classifiers

Challenges

- What information about classifiers should be represented?
- Variable number of actions
- Size of action space increases over time
- Number of classifiers increases over time
- Very large action space after initial interactions.

Tackling challenges

- Features based on active learning methods
  - Representing classifiers
- Featurize state-action pairs
  - Variable number of actions and classifiers
- Sampling a beam of promising queries
  - Large action space

Feature Groups

- Query features - Active learning metrics used to determine whether a query is useful
- Guess features - Features that use the predictions and confidences of classifiers to determine whether a guess will be correct

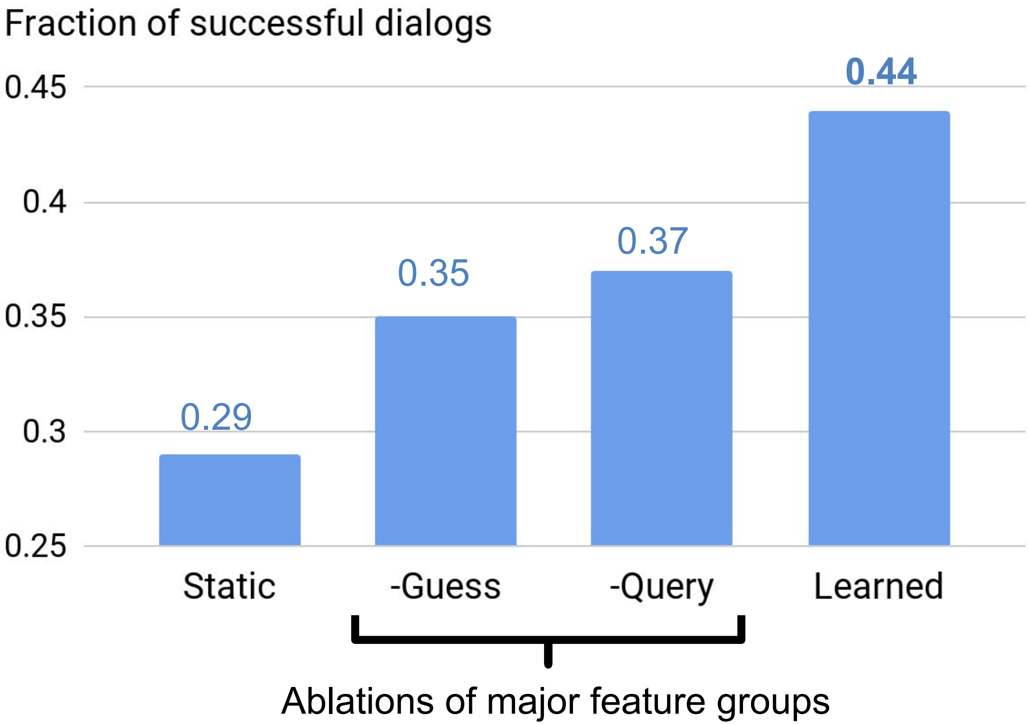
Experiment Setup

- Policy learning using REINFORCE.
- Baseline - A hand-coded dialog policy that asks a fixed number of questions selected using the same sampling distribution.

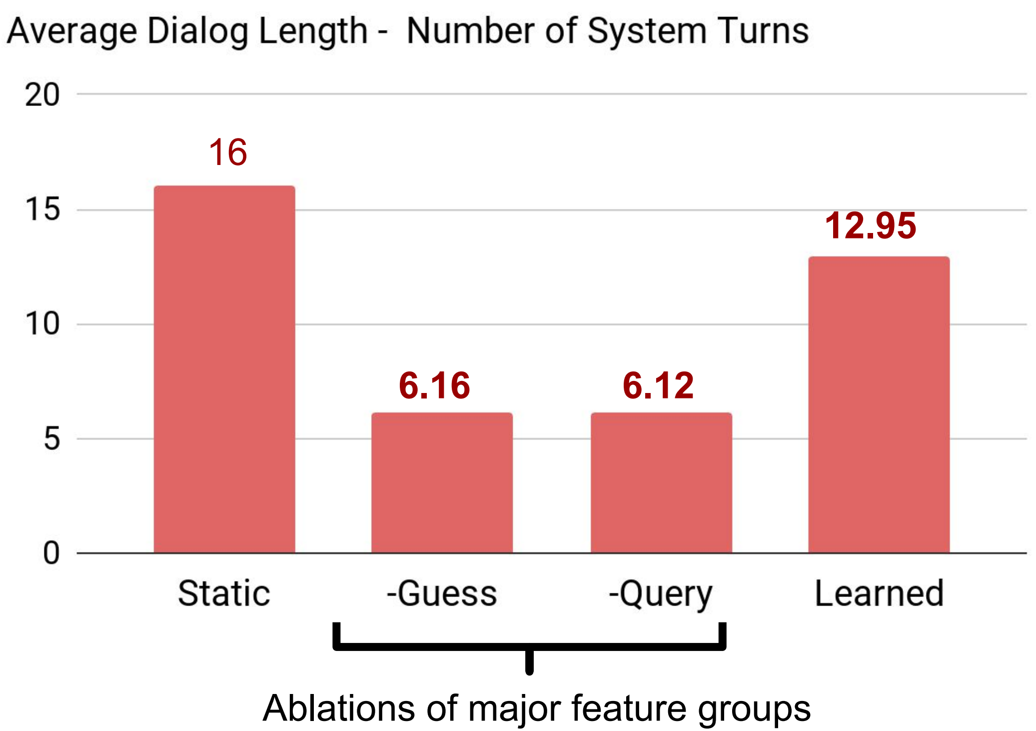
Experiment Phases

- Initialization - Collect experience using the baseline to initialize the policy.
- Training - Improve the policy from on-policy experience.
- Testing - Policy weights are fixed, and we run a new set of interactions, starting with no classifiers, over an independent test set with different predicates.

Results


Fraction of successful dialogs
0.45
0.4
0.35
0.3
0.25
0.29
0.35
0.37
0.44
Static
-Guess
-Query
Learned
Ablations of major feature groups

Results


Average Dialog Length - Number of System Turns
20
15
10
5
0
16
6.16
6.12
12.95
Static
-Guess
-Query
Learned
Ablations of major feature groups

Summary

- We can learn a dialog policy that learns to acquire knowledge of predicates through opportunistic active learning.
- The learned policy is more successful at object retrieval than a static baseline, using fewer dialog turns on average.

Outline

- Background
- Completed Work
  - Integrating Learning of Dialog Strategies and Semantic Parsing (Padmakumar et.al., 2017)
  - Opportunistic Active Learning for Grounding Natural Language Descriptions (Thomason et. al., 2017)
  - Learning a Policy for Opportunistic Active Learning (Padmakumar et. al., 2018)
- Proposed Work
- Conclusion

Outline

- Proposed Work
  - Learning to Ground Natural Language Object Descriptions Using Joint Embeddings
  - Identifying Useful Clarification Questions for Grounding Object Descriptions
  - Learning a Policy for Clarification Questions using Uncertain Models
  - Bonus Contributions

Perceptual Grounding Using Classifiers

Perceptual Grounding

blue mug ➡ $\lambda x.(\mathtt{blue}(x) \wedge \mathtt{mug}(x))$

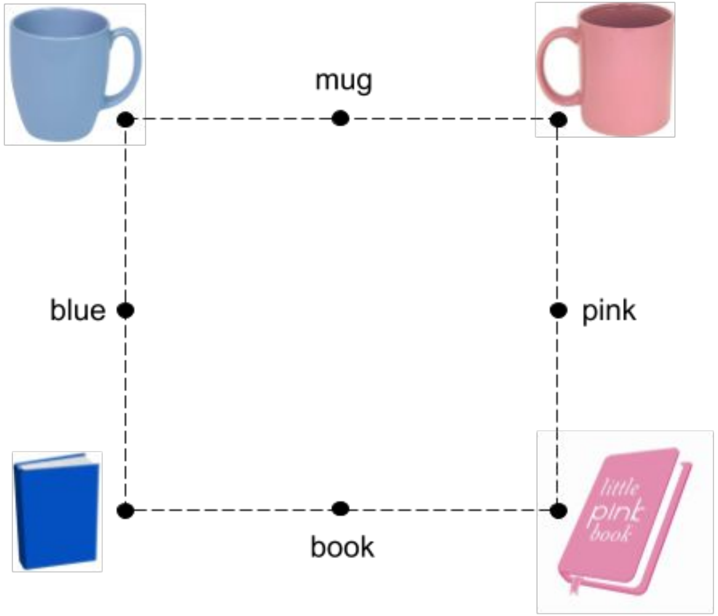
➡ Classifier `blue/not blue` ➡ **`blue`**

➡ Classifier `mug/not mug` ➡ **`mug`**

➡ Classifier `blue/not blue` ➡ **`not blue`**

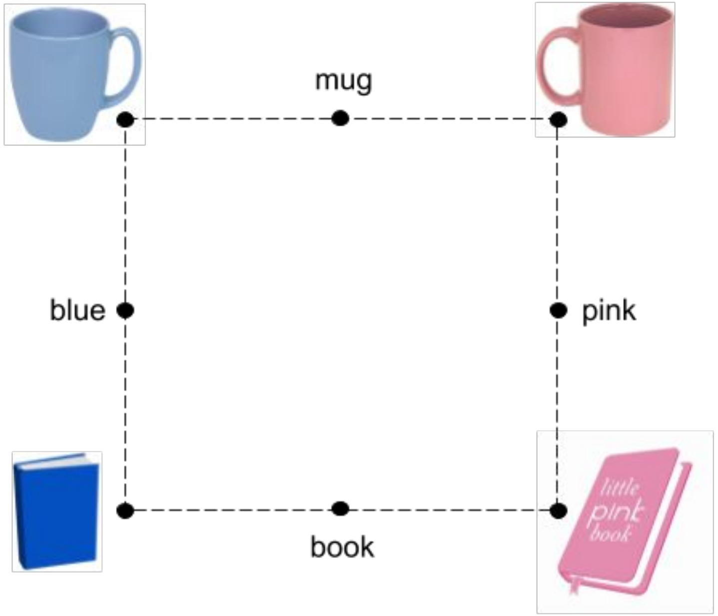
➡ Classifier `mug/not mug` ➡ **`mug`**

Grounding Using a Joint Vector Space



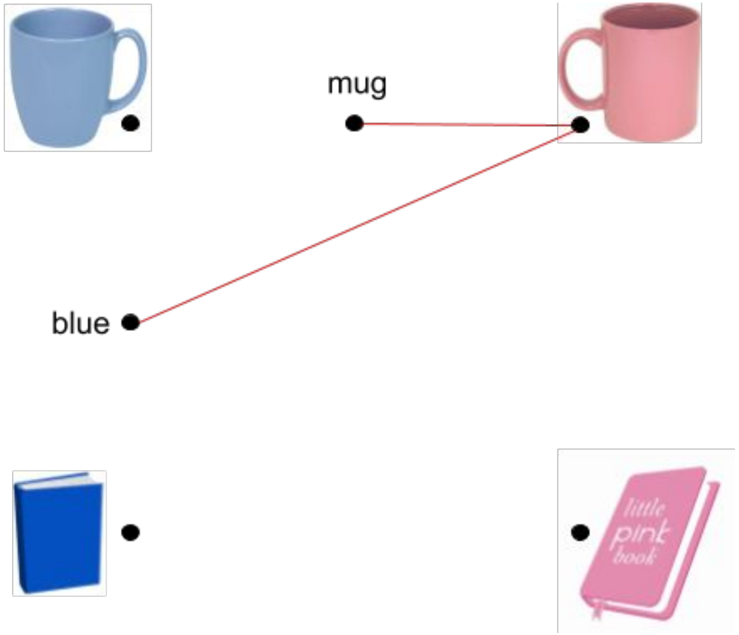
Grounding Using a Joint Vector Space

- Represent words and images as vectors in the same space.
- Words are near images they apply to and vice versa.



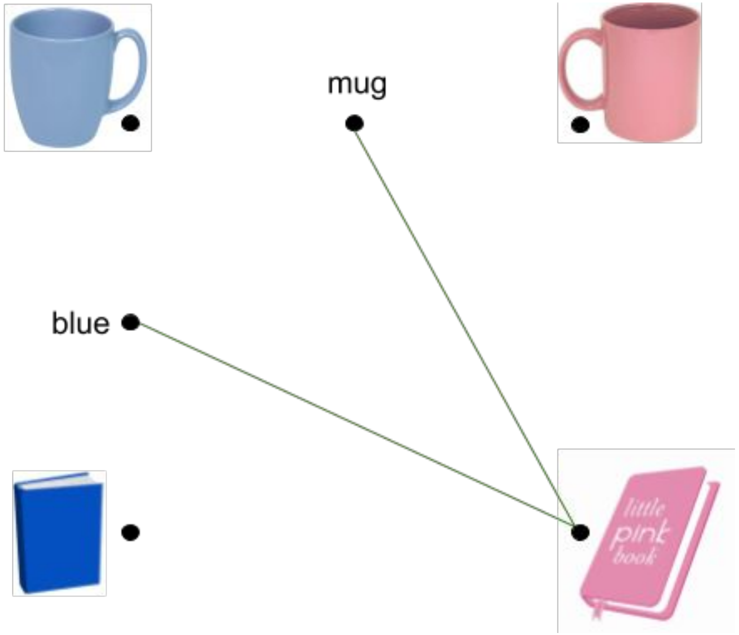
Grounding Using a Joint Vector Space

To ground a description, such as “blue mug”, find the image which minimizes the sum of distances to the words.



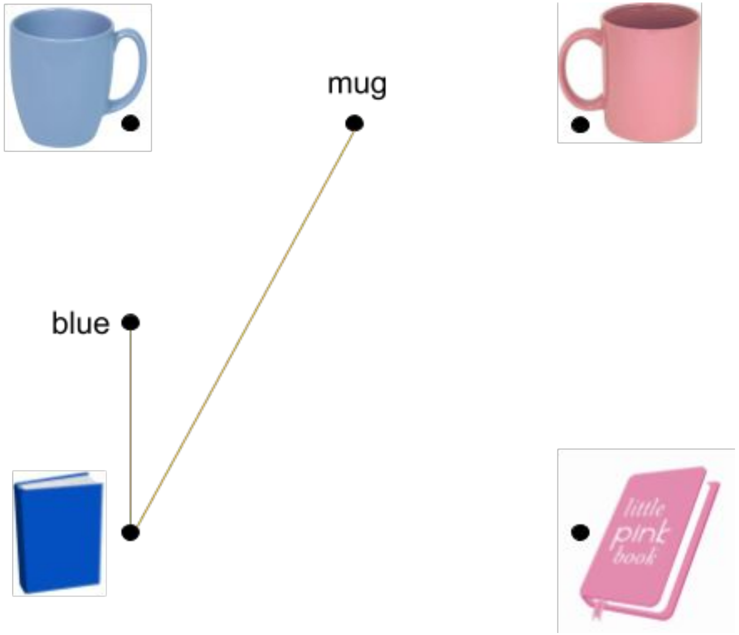
Grounding Using a Joint Vector Space

To ground a description, such as “blue mug”, find the image which minimizes the sum of distances to the words.



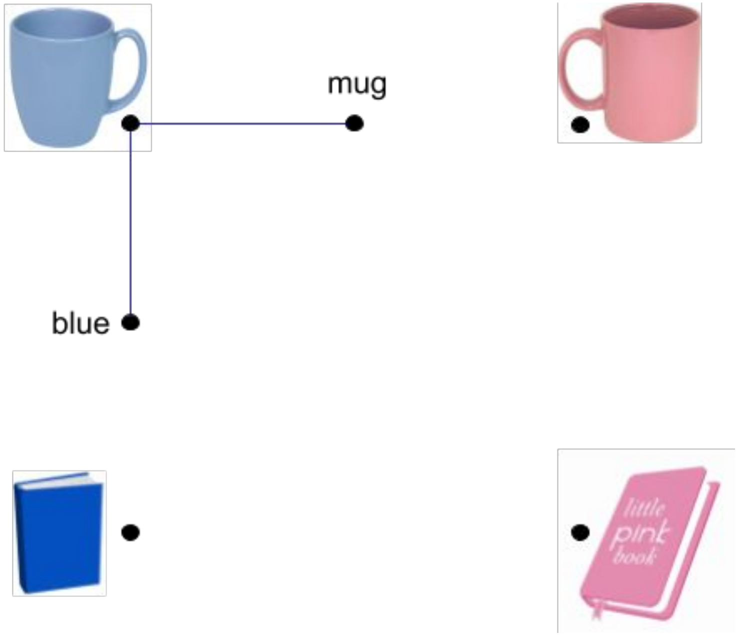
Grounding Using a Joint Vector Space

To ground a description, such as "blue mug", find the image which minimizes the sum of distances to the words.



Grounding Using a Joint Vector Space

To ground a description, such as "blue mug", find the image which minimizes the sum of distances to the words.

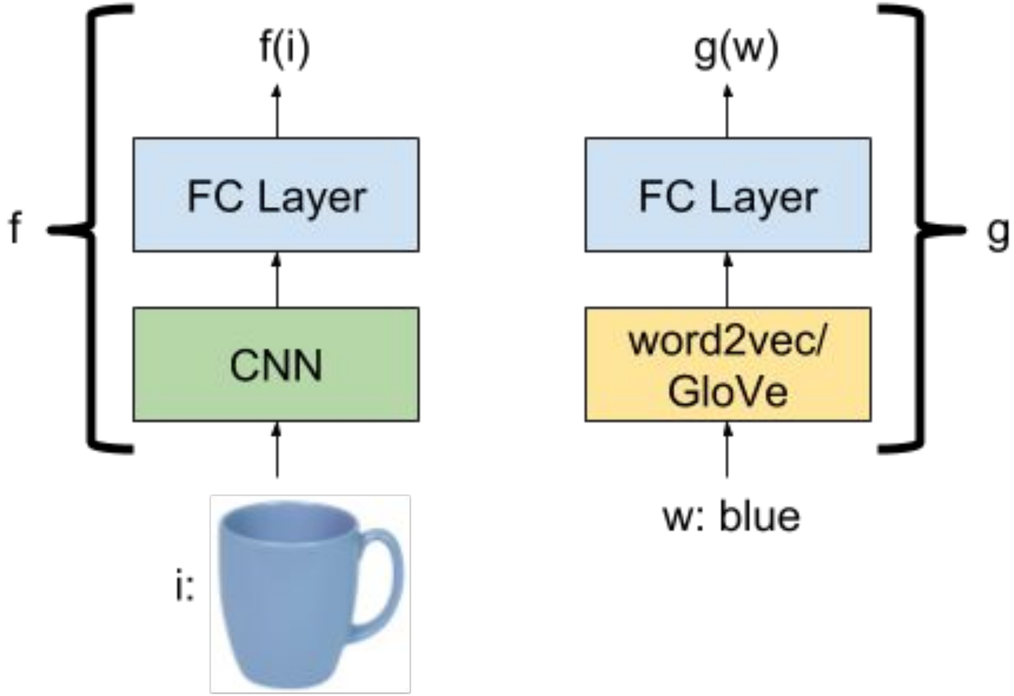


Grounding Using a Joint Vector Space

Related prior work

- Word vectors in learned joint spaces are more useful for many tasks, eg: semantic relatedness [Lazaridou et. al., 2015]
- Neural networks that score an image-description pair perform well at grounding but use sentence embeddings [Hu et. al. 2016, Xiao et. al. 2017].
- We expect that words would generalize better than phrases/ sentences.

Learning the Joint Space



Learning the Joint Space

d(f(), g(blue)) ≤ d(f(), g(blue))

d(f(), g(blue)) ≤ d(f(), g(pink))

Constraints captured using a ranking loss

Outline

- Proposed Work
  - Learning to Ground Natural Language Object Descriptions Using Joint Embeddings
  - Identifying Useful Clarification Questions for Grounding Object Descriptions
  - Learning a Policy for Clarification Questions using Uncertain Models
  - Bonus Contributions

Identifying Useful Clarification Questions for Grounding Object Descriptions

Bring the blue mug from Alice's office

What should I bring?

The blue coffee mug

What should I bring?

Identifying Useful Clarification Questions for Grounding Object Descriptions

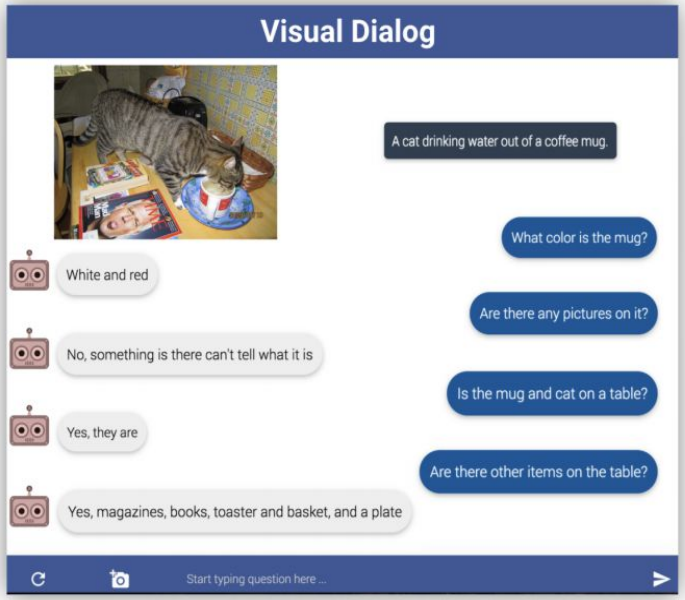
Bring the blue mug from Alice's office

Is this the object I should bring?

No

Recent Related Work

[Das, et. al., 2017]



[De Vries et. al., 2017]

Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Identifying Useful Clarification Questions for Grounding Object Descriptions

- Clarification questions that help narrow down an object being referred to.
- More specific than a new description.
- More general than showing each possible object.
- Provide ground truth answers to questions at training time to learn human semantics.

Attribute Based Queries

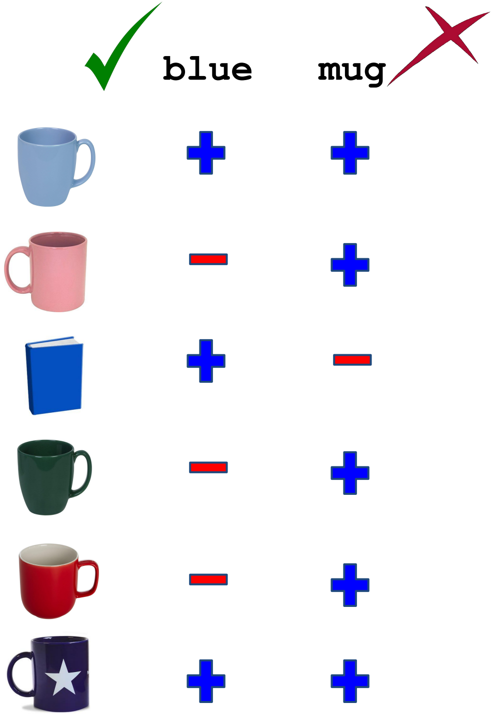
Bring the blue mug from Alice's office

Is the object I should bring a cup?

Yes

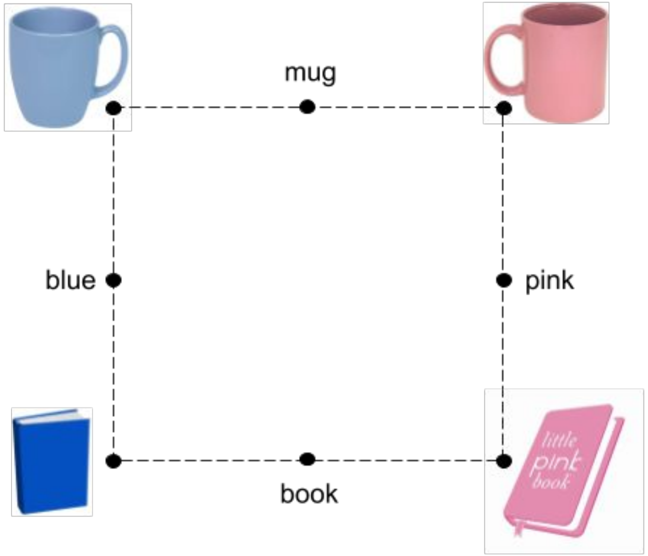
Choosing a Good Query

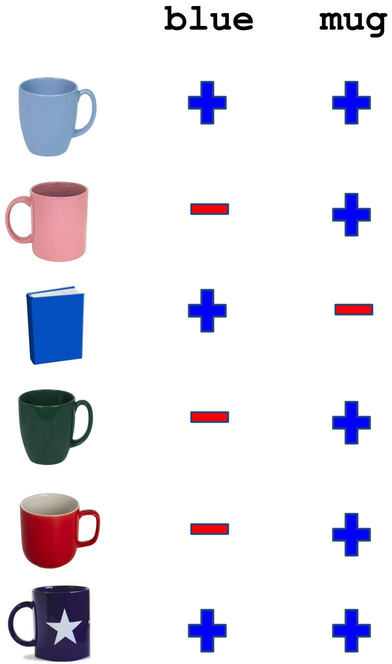
- Query that is most likely to reduce the search space.
- Choose the attribute with respect to which the dataset has highest entropy



Challenge

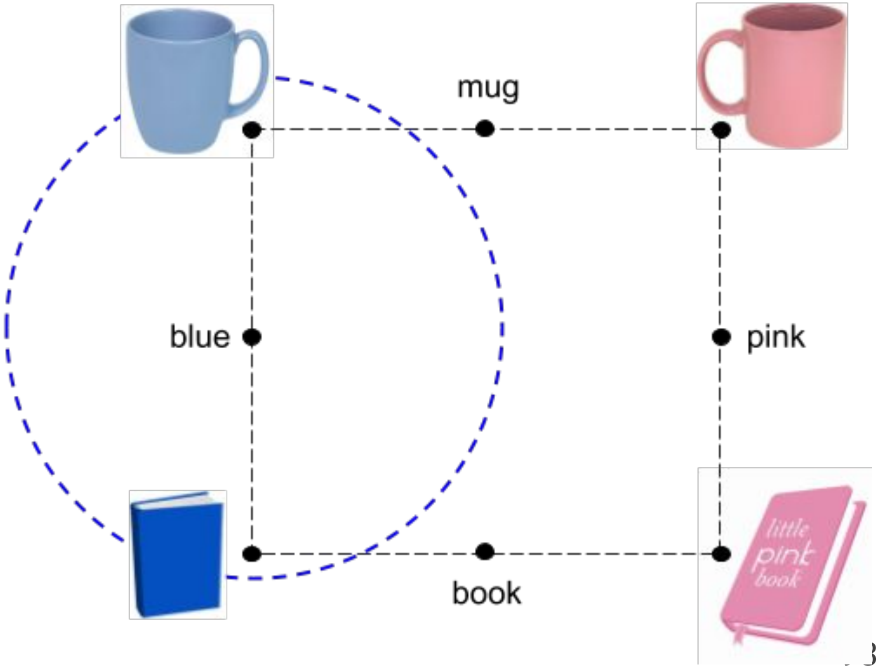
In a joint embedding space how do you determine whether an attribute is applicable?





Possible solutions

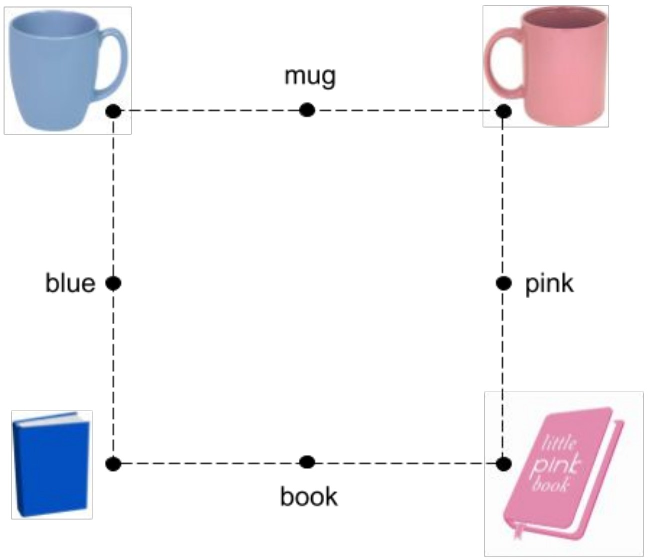
- Distance threshold, clustering to get classifier-like predictions.
- Might be possible to formulate an optimization problem using distances.

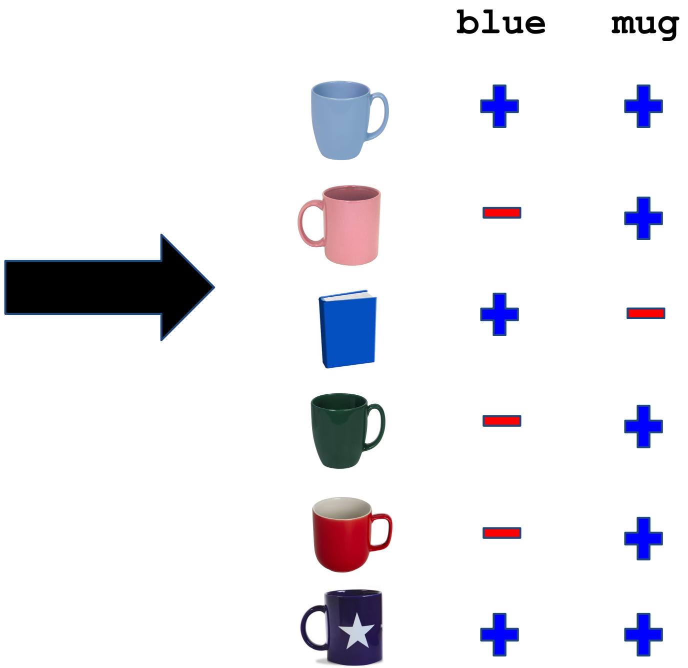


Outline

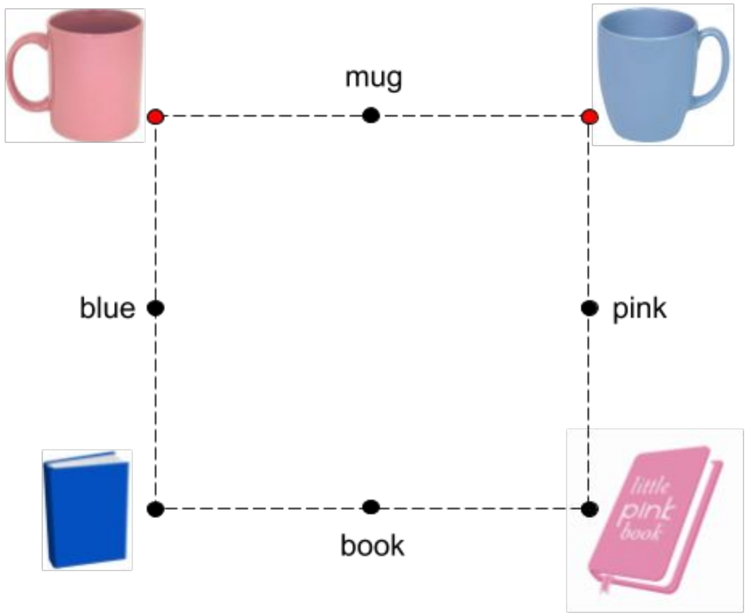
- Proposed Work
  - Learning to Ground Natural Language Object Descriptions Using Joint Embeddings
  - Identifying Useful Clarification Questions for Grounding Object Descriptions
  - Learning a Policy for Clarification Questions using Uncertain Models
  - Bonus Contributions

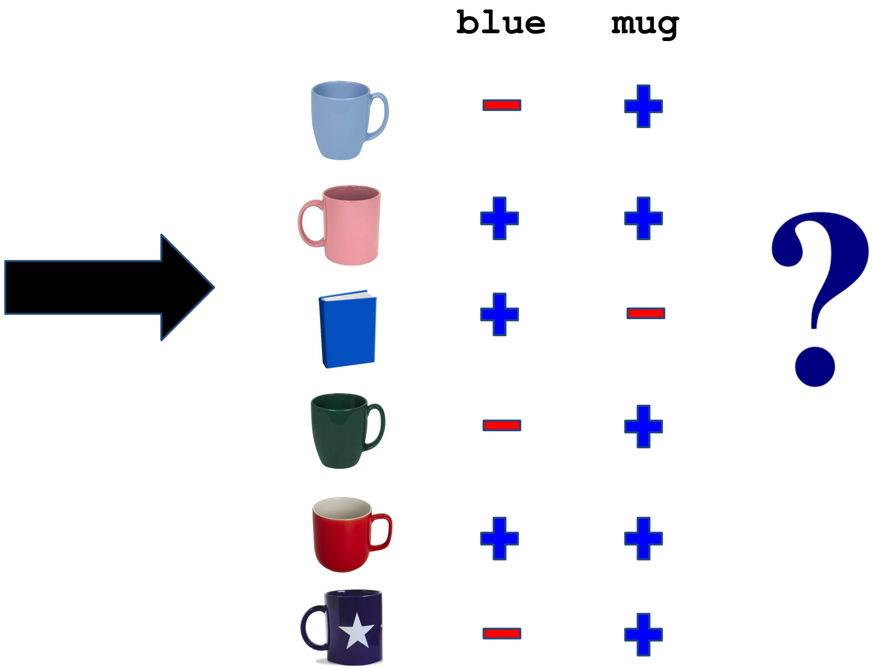
Learning a Policy for Clarification Questions using Uncertain Models





Learning a Policy for Clarification Questions using Uncertain Models



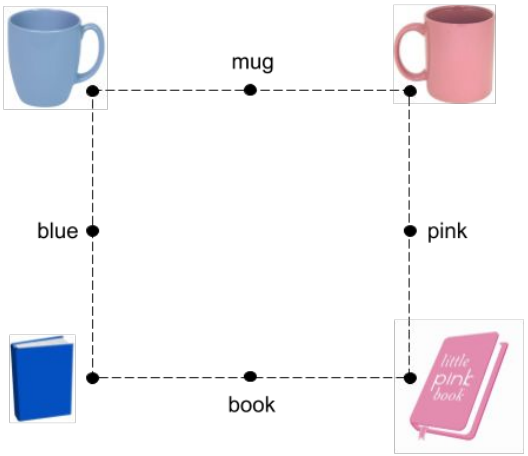


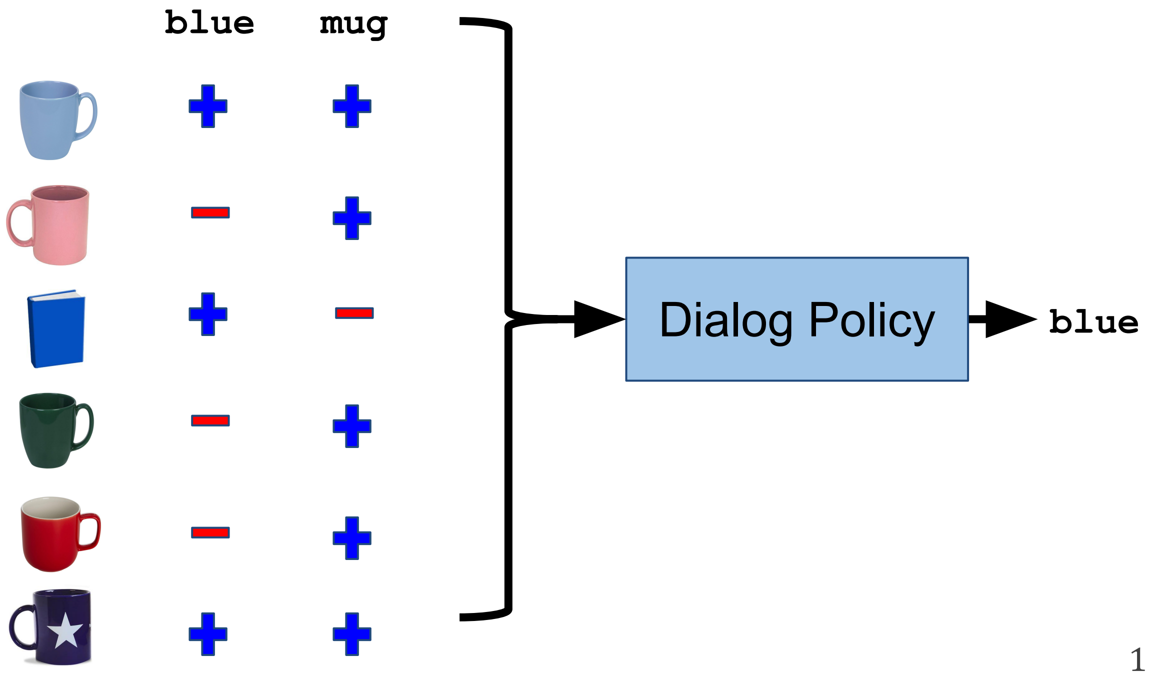
Learning a Policy for Clarification Questions using Uncertain Models

- Proposed method for identifying good queries assumes that the learned space is “good”.
- If predictions for some attribute are especially unreliable, it might be preferable to choose another attribute that is less informative but more reliable.

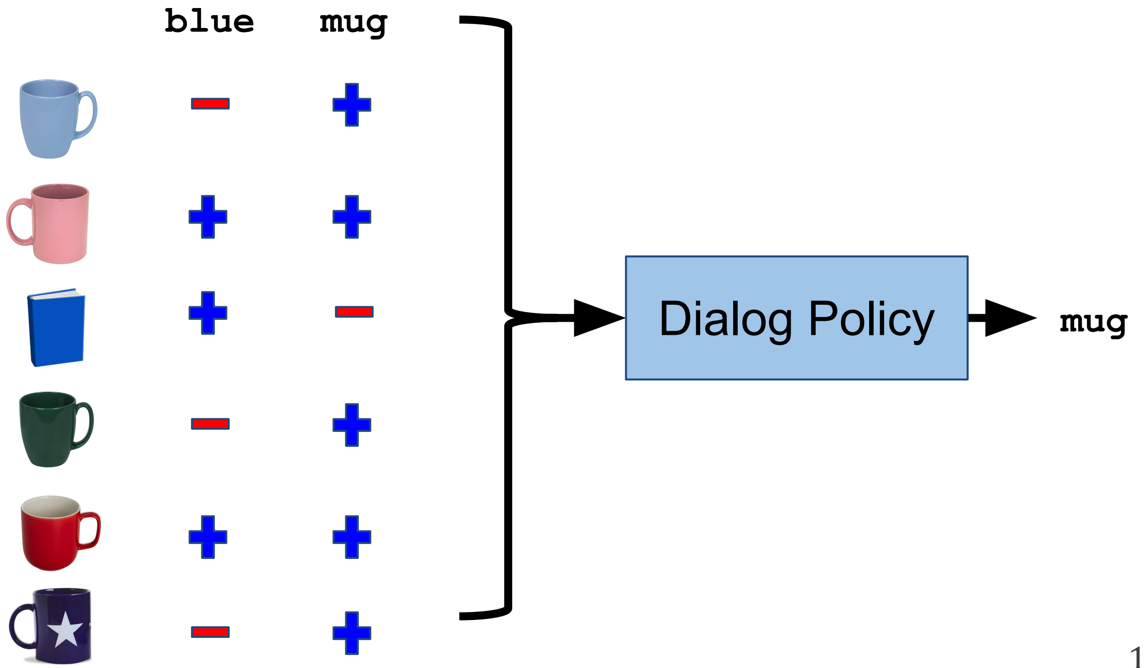
Learning a Policy for Clarification Questions using Uncertain Models

Bring the blue mug from Alice's office





Learning a Policy for Clarification Questions using Uncertain Models



Challenge

- The policy needs features that measure "how good" the space is.
  - Number of training examples
  - How often are the space constraints satisfied?

d(f(), g(blue)) ≤ d(f(), g(blue))

d(f(), g(blue)) ≤ d(f(), g(pink))

?

Outline

- Proposed Work
  - Learning to Ground Natural Language Object Descriptions Using Joint Embeddings
  - Identifying Useful Clarification Questions for Grounding Object Descriptions
  - Learning a Policy for Clarification Questions using Uncertain Models
  - Bonus Contributions

Incorporating Linguistic and Visual Context

water **glass**

wine **glass**

looking **glass**

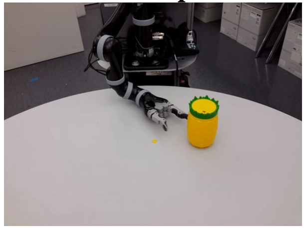
glass swan

the big bottle

the small bottle

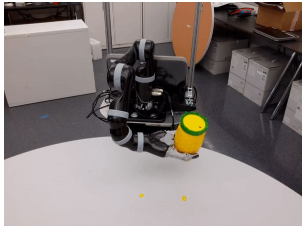
Using Multimodal Object Representations

Grasp



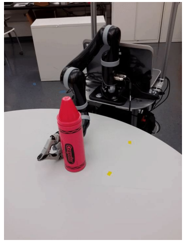
Lift

Lower



Drop

Press



Push

Outline

- Background
- Completed Work
  - Integrating Learning of Dialog Strategies and Semantic Parsing (Padmakumar et.al., 2017)
  - Opportunistic Active Learning for Grounding Natural Language Descriptions (Thomason et. al., 2017)
  - Learning a Policy for Opportunistic Active Learning (Padmakumar et. al., 2018)
- Proposed Work
- Conclusion

Improving Natural Language Understanding Through Dialog

Bring the blue mug from Alice's office

Where should I bring a blue mug from?

Alice Ashcraft's office

I should bring a blue mug from 3502?

Yes

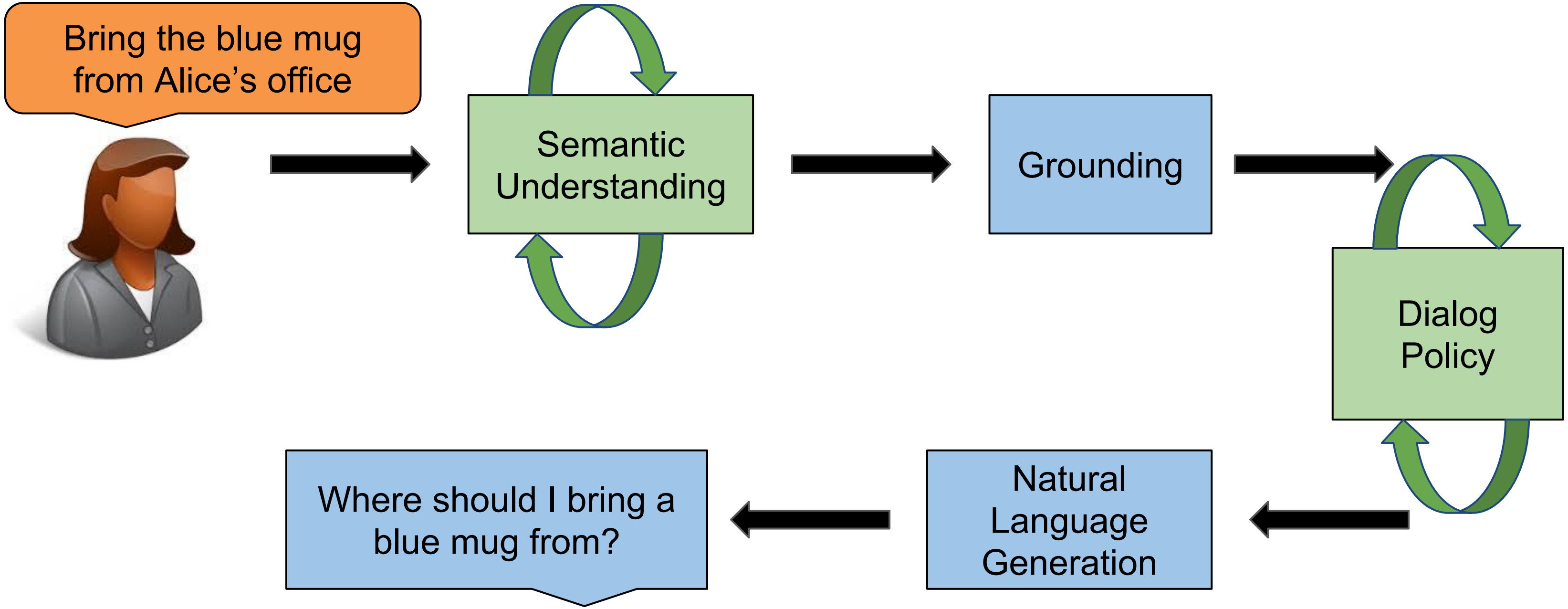
Improving Natural Language Understanding Through Dialog

Bring the blue mug from Alice's office

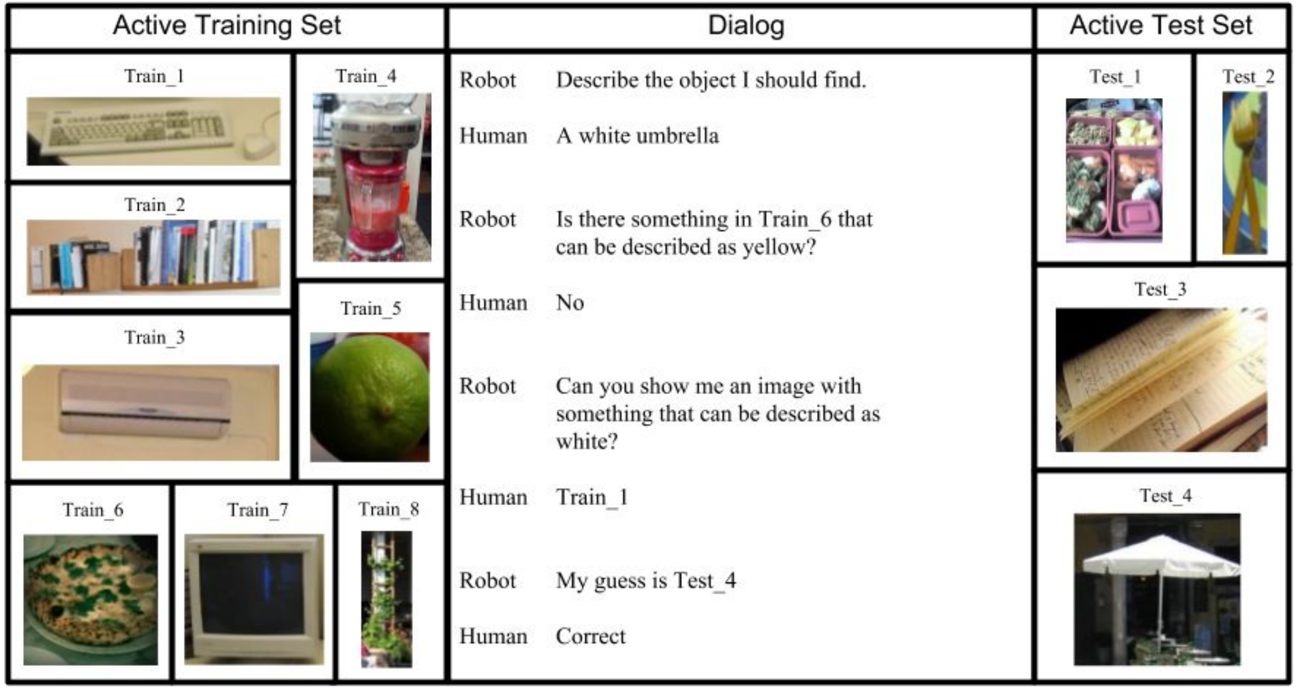
Would you use the word "blue" to refer to this object?

Yes

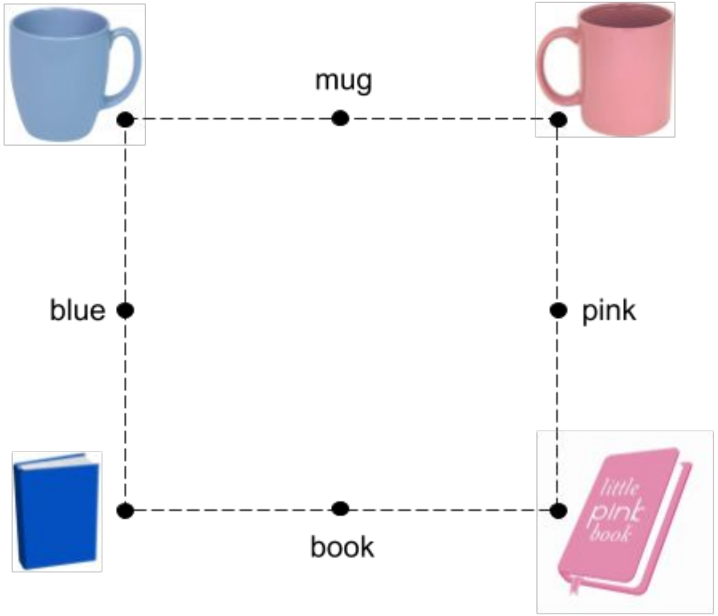
Joint Parser and Policy Learning



Policy Learning for Opportunistic Active Learning



Improved Perceptual Grounding Model



Clarification Questions for Object Descriptions

Bring the blue mug from Alice's office

Is the object I should bring a cup?

Yes

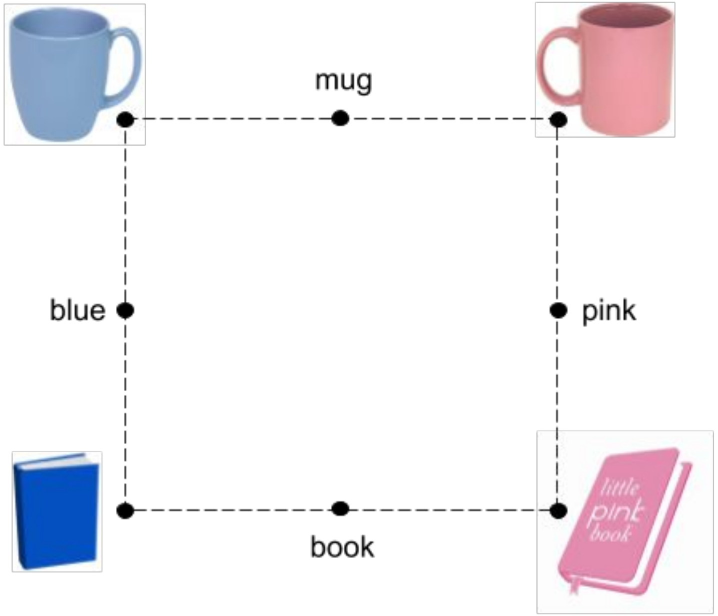
Improved Models and Queries for Grounded Human-Robot Dialog

Aishwarya Padmakumar

Doctoral Dissertation Proposal

Incorporating Context

- Visual Context
  - Representations of other objects
  - Representation of the entire scene and the object's bounding box
- Linguistic Context - ELMo embeddings



Learning Joint Embeddings with Multimodal Object Representations

- Not all modalities are equally informative for each object-word pair.
- Not all modalities may be available for each object.
- Project features of each modality to the same space and combine during grounding.

Computing Distance

- Average distance of the word to object representation in all modalities.
- Distance of the word to nearest object representation - allows only one modality to be relevant.