COMP 790.139 (Fall 2017)
Natural Language Processing

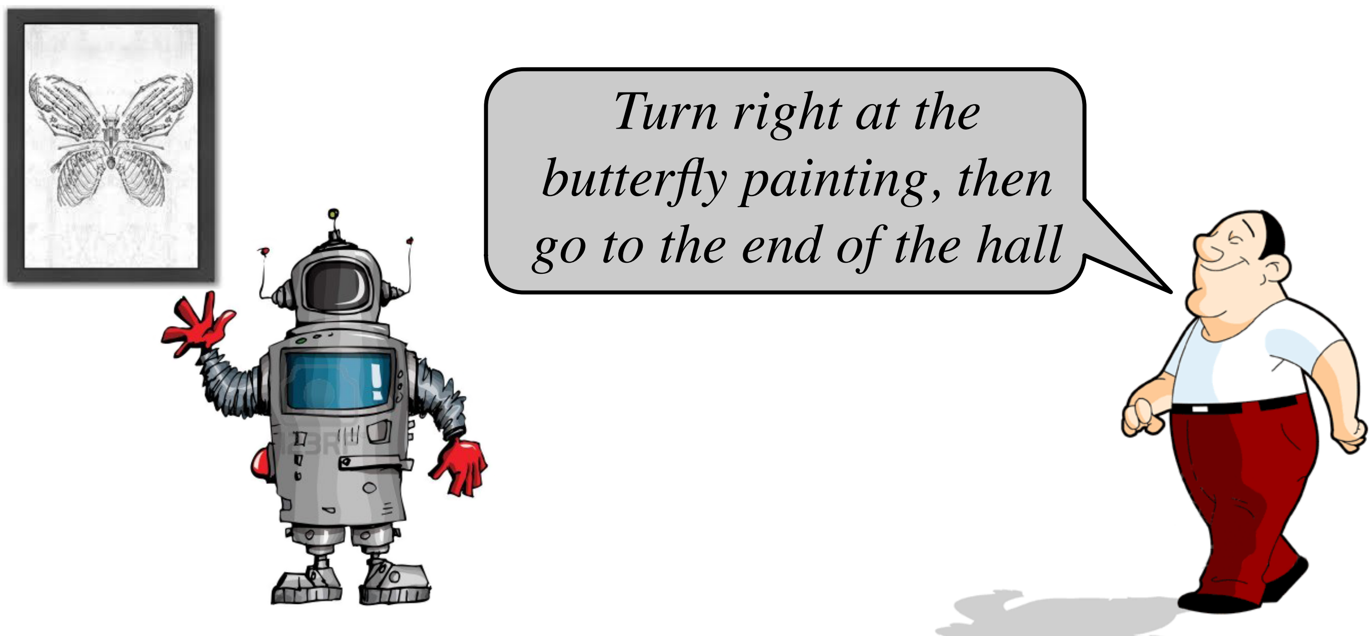
Language+Robotics

THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

Mohit Bansal

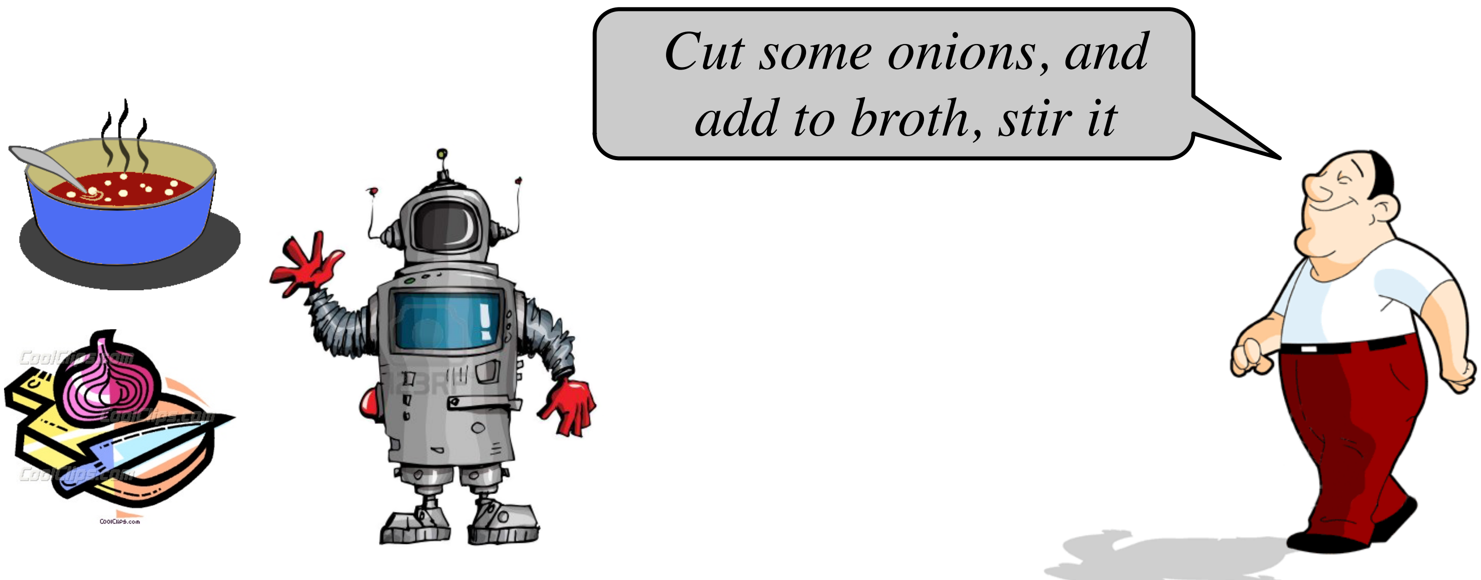
NLP for Actions/Robotics

- Task-based instructions, e.g., navigation, grasping, manipulation, skill learning



NLP for Actions/Robotics

- Task-based instructions, e.g., navigation, grasping, manipulation, skill learning



Navigation Instruction Following

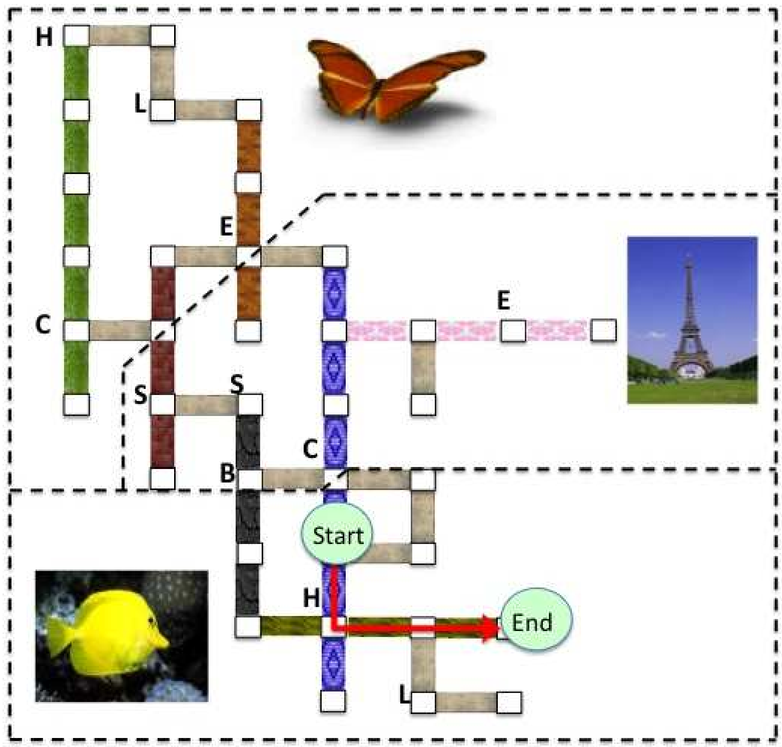
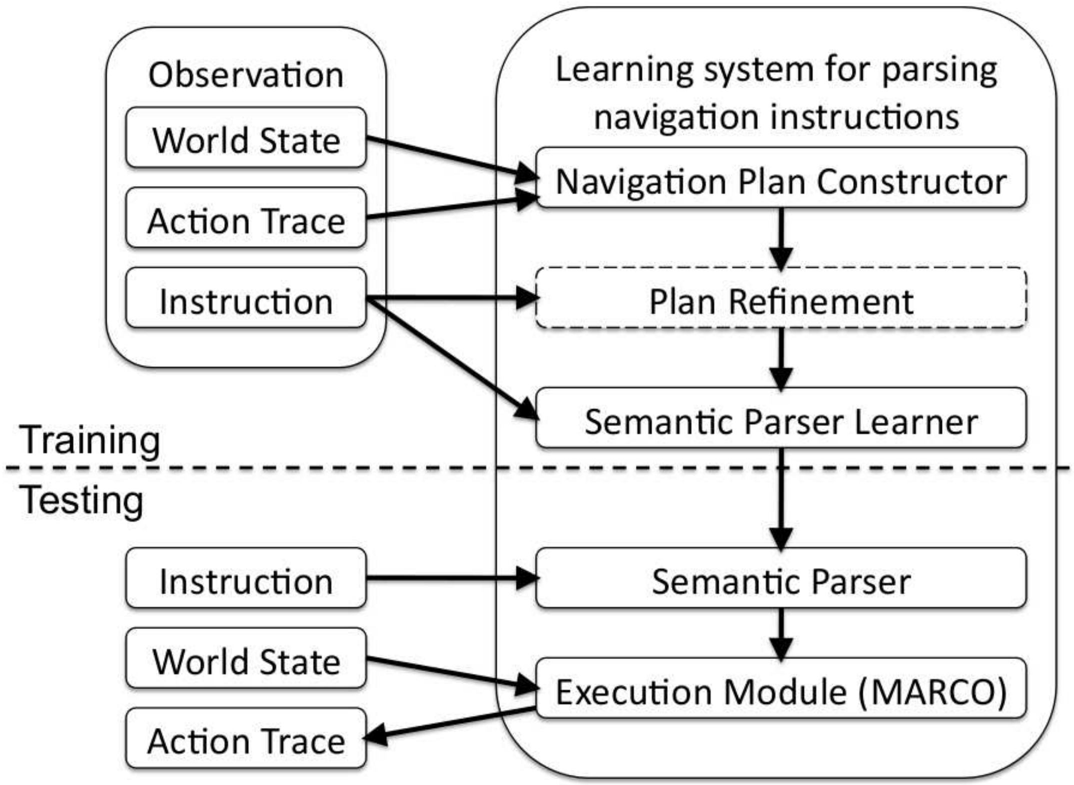


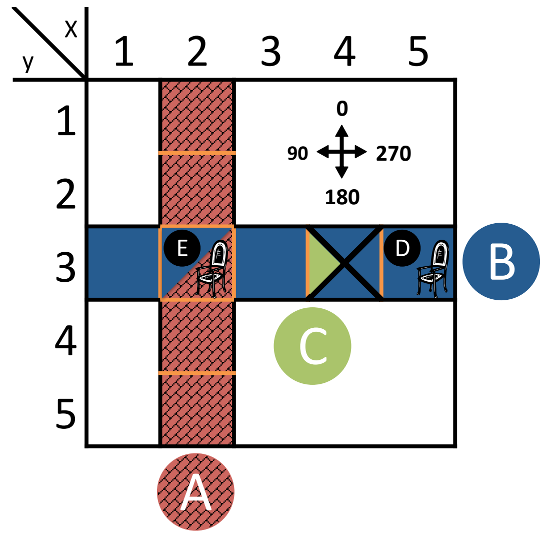
Figure 1: This is an example of a route in our virtual world. The world consists of interconnecting hallways with varying floor tiles and paintings on the wall (butterfly, fish, or Eiffel Tower.) Letters indicate objects (e.g. 'C' is a chair) at a location.

Navigation Instruction Following



Learning to interpret natural language navigation instructions from observations. Chen and Mooney. AAAI 2011.

Navigation Instruction Following



{D E}	(a) chair $\lambda x.chair(x)$
{A B}	(b) hall $\lambda x.hall(x)$
E	(c) the chair $\iota x.chair(x)$
C	(d) you you
{B}	(e) blue hall $\lambda x.hall(x) \wedge blue(x)$
{E}	(f) chair in the intersection $\lambda x.chair(x) \wedge$ $intersect(\iota y.junction(y), x)$
{A B E}	(g) in front of you $\lambda x.in\_front\_of(you, x)$

Figure 2: Schematic diagram of a map environment and example of semantics of spatial phrases.

Weakly supervised learning of semantic parsers for mapping instructions to actions. Artzi and Zettlemoyer. TACL 2013.

Navigation Instruction Following

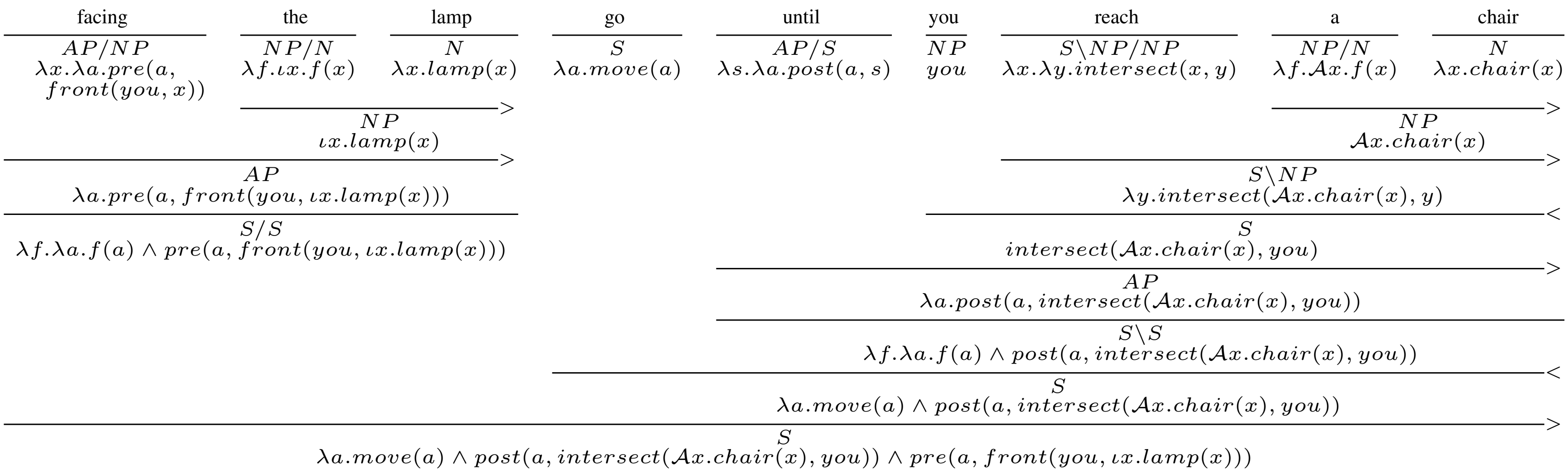
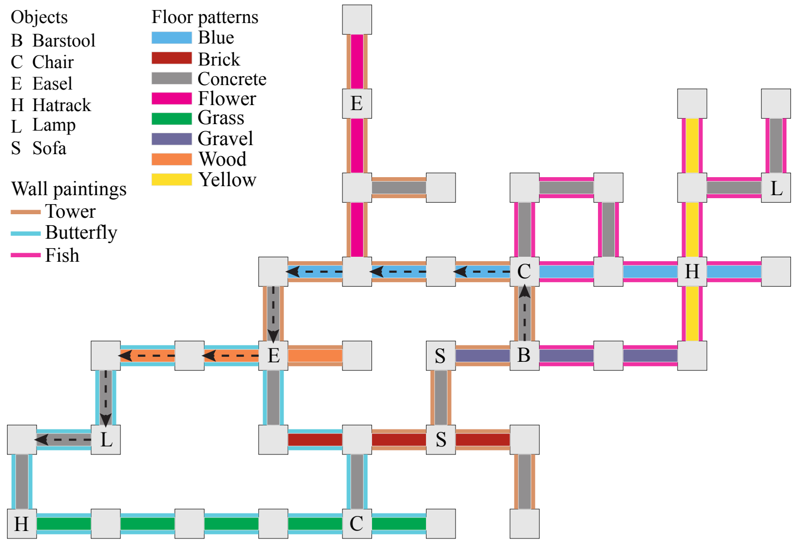


Figure 4: A CCG parse showing adverbial phrases and topicalization.

Navigation Instruction Following



> Place your back against the wall of the "T" intersection. Go forward one segment to the intersection with the blue-tiled hall. This interesction [sic] contains a chair. Turn left. Go forward to the end of the hall. Turn left. Go forward one segment to the intersection with the wooden-floored hall. This intersection conatains [sic] an easel. Turn right. Go forward two segments to the end of the hall. Turn left. Go forward one segment to the intersection containing the lamp. Turn right. Go forward one segment to the empty corner.

Figure 1: An example of a route instruction-path pair in one of the virtual worlds from MacMahon, Stankiewicz, and Kuipers (2006) with colors that indicate floor patterns and wall paintings, and letters that indicate different objects. Our method successfully infers the correct path for this instruction.

Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences. Mei, Bansal, and Walter. AAAI 2016

Navigation Instruction Following

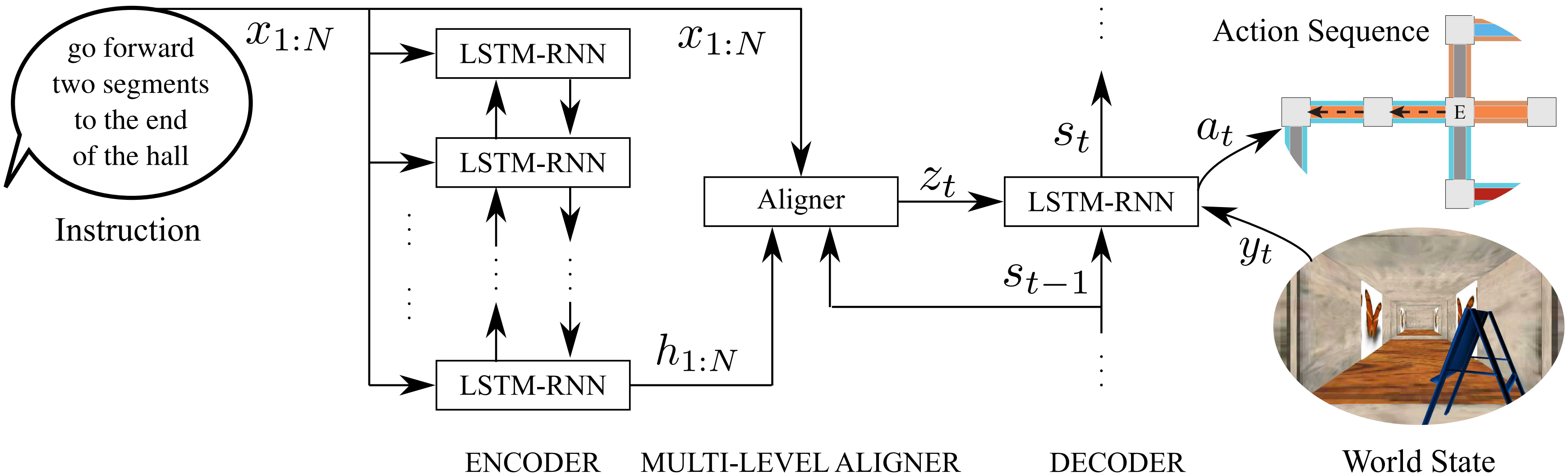


Figure 2: Our encoder-aligner-decoder model with multi-level alignment

Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences. Mei, Bansal, and Walter. AAAI 2016

Navigation Instruction Following

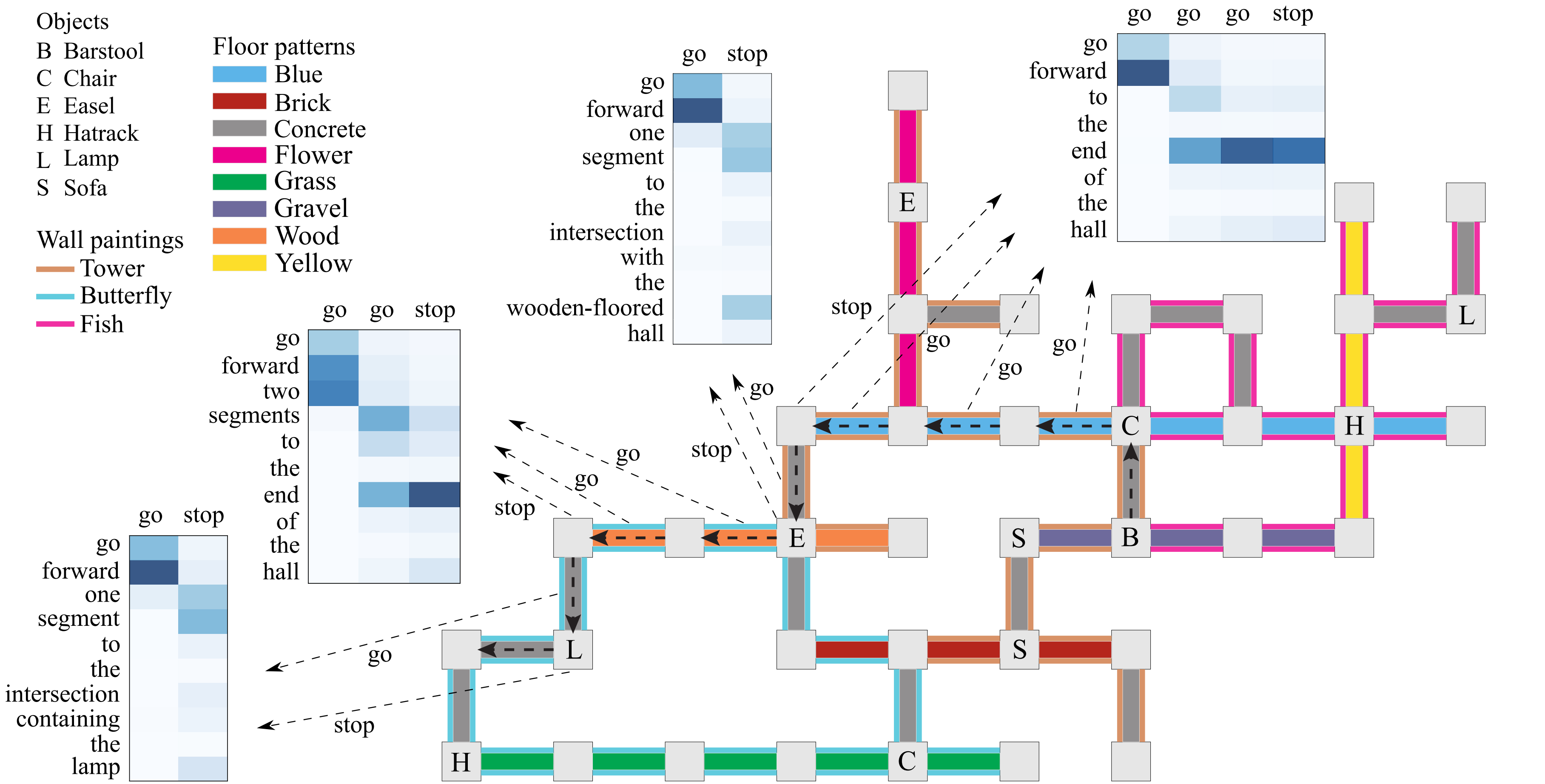


Figure 4: Visualization of the alignment between words to actions in a map for a multi-sentence instruction.

Navigation Instruction Following

Table 1: Overall accuracy (state-of-the-art in bold)

Method	Single-sent	Multi-sent
Chen and Mooney (2011)	54.40	16.18
Chen (2012)	57.28	19.18
Kim and Mooney (2012)	57.22	20.17
Kim and Mooney (2013)	62.81	26.57
Artzi and Zettlemoyer (2013)	65.28	31.93
Artzi, Das, and Petrov (2014)	64.36	**35.44**
Andreas and Klein (2015)	59.60	–
Our model (vDev)	69.98	26.07
Our model (vTest)	**71.05**	30.34

Navigation Instruction Following

Table 2: Model components ablations

	Full Model	High-level Aligner	No Aligner	Unidirectional	No Encoder
Single-sentence	69.98	68.09	68.05	67.44	61.63
Multi-sentence	26.07	24.79	25.04	24.50	16.67

Table 3: Accuracy as a function of distance from destination

Distance (d)	0	1	2	3
Single-sentence	71.73	86.62	92.86	95.74
Multi-sentence	26.07	42.88	59.54	72.08

Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences. Mei, Bansal, and Walter. AAAI 2016

Navigation + Manipulation Instructions

Commands from the corpus
- Go to the first crate on the left and pick it up.
- Pick up the pallet of boxes in the middle and place them on the trailer to the left.
- Go forward and drop the pallets to the right of the first set of tires.
- Pick up the tire pallet off the truck and set it down

(a) Robotic forklift (b) Sample commands

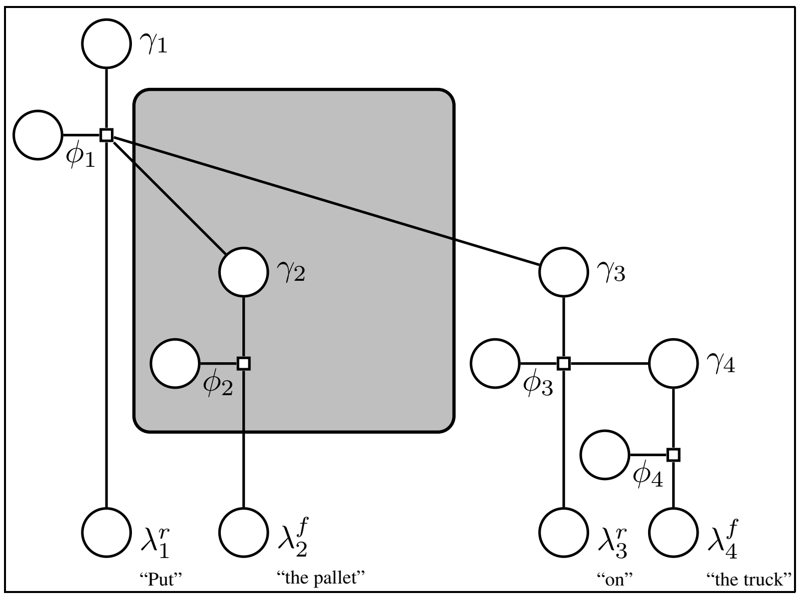
Figure 1: A target robotic platform for mobile manipulation and navigation (Teller et al., 2010), and sample commands from the domain, created by untrained human annotators. Our system can successfully follow these commands.

Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. Tellex, Kollar, Dickerson, Walter, Banerjee, Teller, and Roy. AAAI 2011.

Navigation + Manipulation Instructions

$EVENT_1(r = \texttt{Put},$
$\quad l = OBJ_2(f = \texttt{the pallet}),$
$\quad l2 = PLACE_3(r = \texttt{on},$
$\qquad l = OBJ_4(f = \texttt{the truck})))$

(a) SDC tree

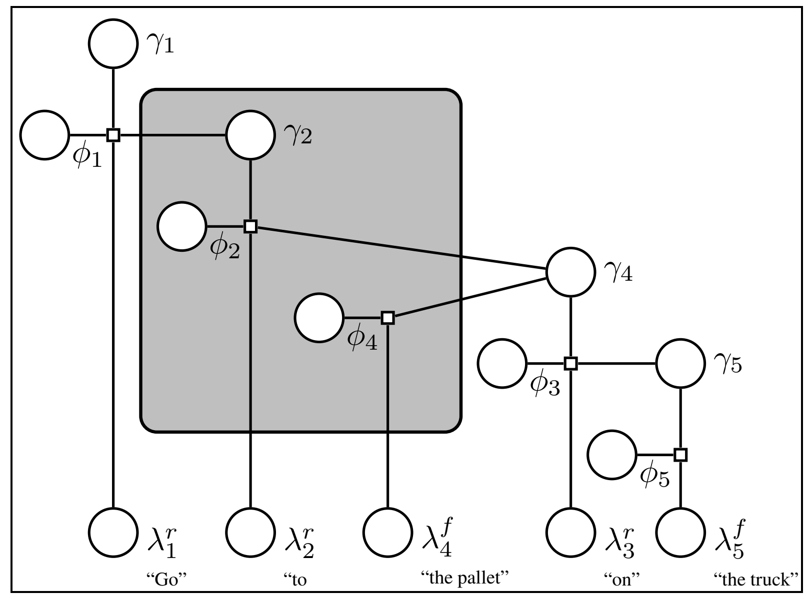


(b) Induced model

Figure 2: (a) SDC tree for "Put the pallet on the truck." (b) Induced graphical model and factorization.

$EVENT_1(r = \texttt{Go}$
$\quad l = PATH_2(r = \texttt{to},$
$\qquad l = OBJ_3(f = OBJ_4(f = \texttt{the pallet}),$
$\qquad\quad r = \texttt{on},$
$\qquad\quad l = OBJ_5(f = \texttt{the truck}))))$

(a) SDC tree

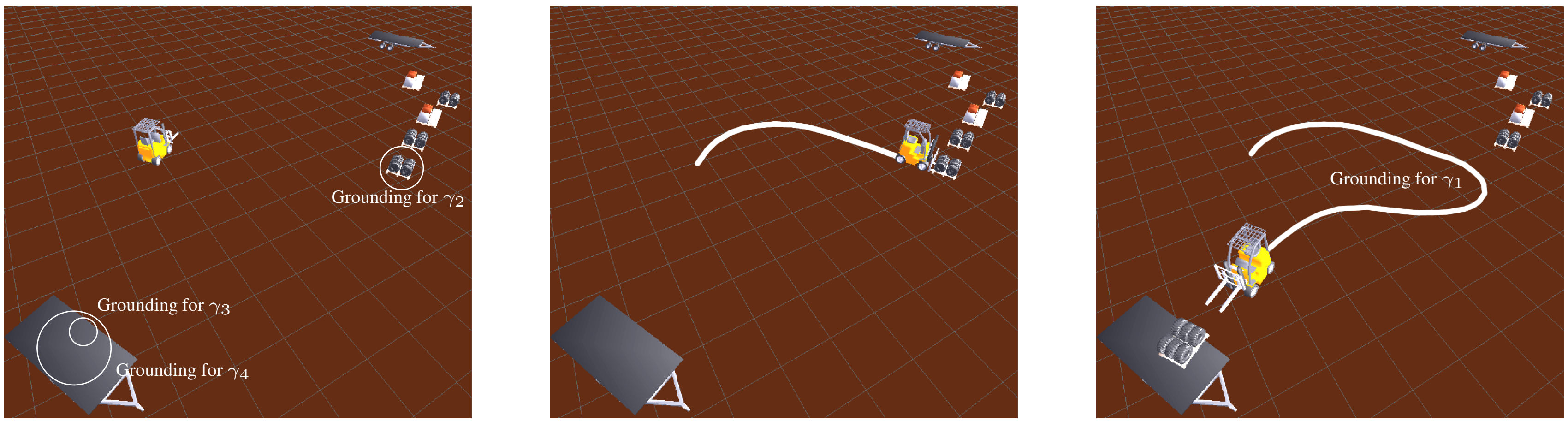


(b) Induced model

Figure 3: (a) SDC tree for "Go to the pallet on the truck." (b) A different induced factor graph from Figure 2. Structural differences between the two models are highlighted in gray.

Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. Tellex, Kollar, Dickerson, Walter, Banerjee, Teller, and Roy. AAAI 2011.

Navigation + Manipulation Instructions



(a) Object groundings (b) Pick up the pallet (c) Put it on the truck

Figure 4: A sequence of the actions that the forklift takes in response to the command, "Put the tire pallet on the truck." (a) The search grounds objects and places in the world based on their initial positions. (b) The forklift executes the first action, picking up the pallet. (c) The forklift puts the pallet on the trailer.

Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. Tellex, Kollar, Dickerson, Walter, Banerjee, Teller, and Roy. AAAI 2011.

Navigation + Manipulation Instructions

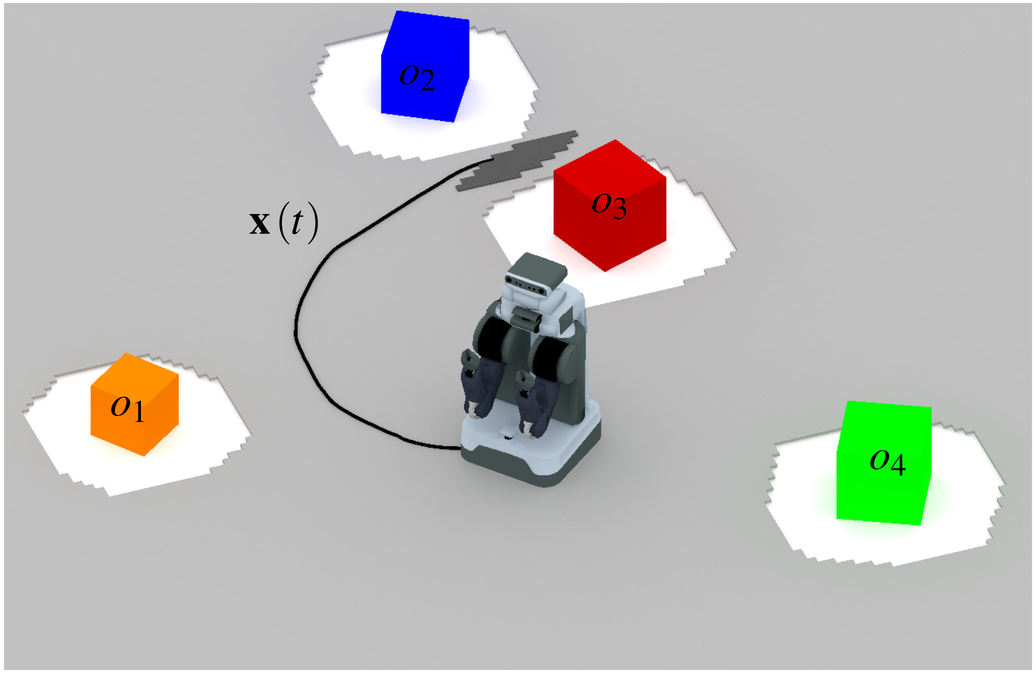


Fig. 1. An illustration of the robot trajectory $\mathbf{x}(t)$ generated from planning constraints that were inferred from the natural language instruction "move near the red box and the blue crate" using the Distributed Correspondence Graph (DCG) model. The dark gray, light gray, and white regions represent the goal states, admissible states, and inadmissible states respectively. The variables $o_1 \ldots o_4$ identify the four objects in the environment model.

A Natural Language Planner Interface for Mobile Manipulators. Howard, Tellex, and Roy. ICRA 2014.

Navigation + Manipulation Instructions

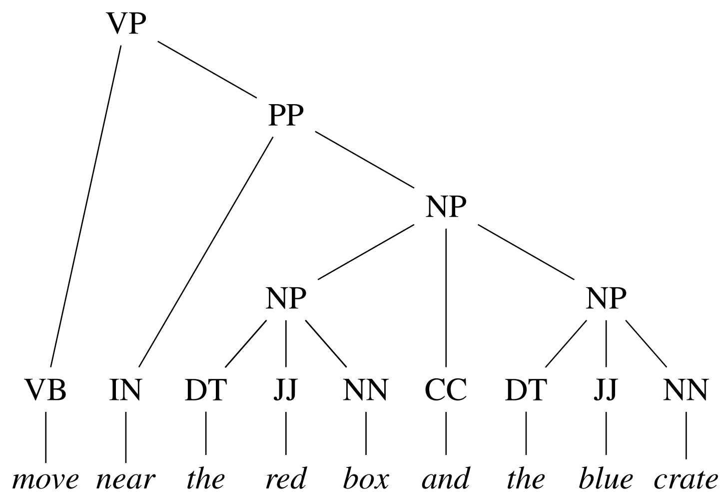


Fig. 2. A parse tree for the sentence "move near the red box and the blue crate". Part-of-speech tags in the parse tree are from the Penn Treebank [2].

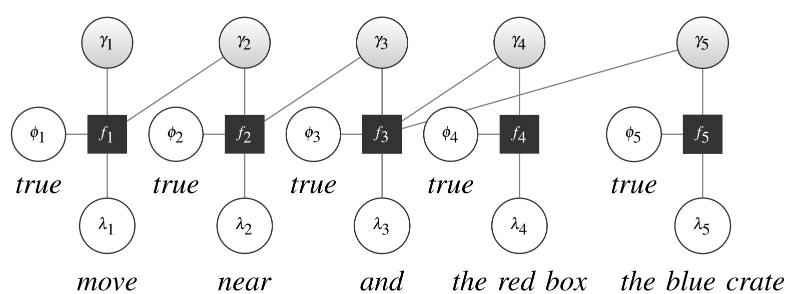
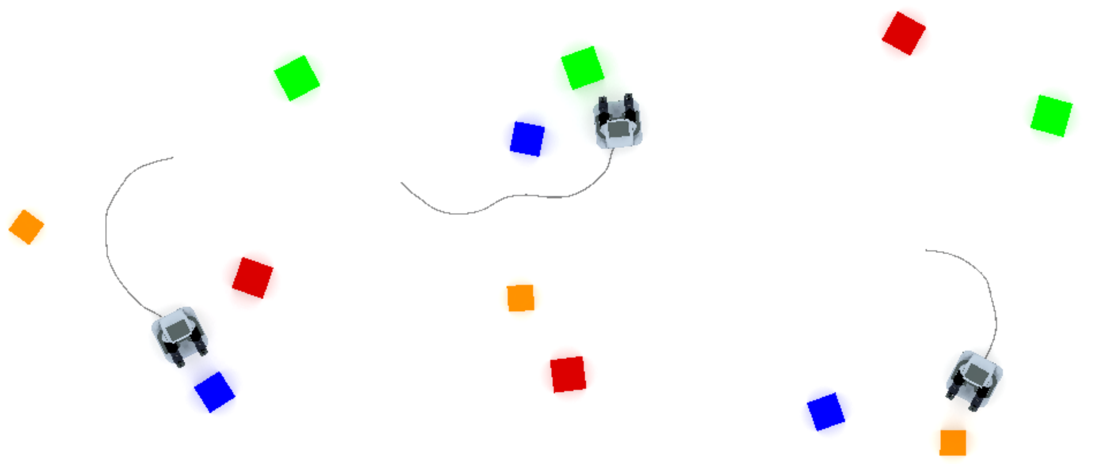


Fig. 3. The factor graph resulting from the parse tree in Figure 2, used by the G^3 algorithm to infer the groundings of the instructions. Each linguistic is grounded to a object, location, or action through a factor that incorporates the grounding of its children. Black boxes, white spheres, and gray sphere are factors, known random variables, and unknown random variables respectively.

A Natural Language Planner Interface for Mobile Manipulators. Howard, Tellex, and Roy. ICRA 2014.

Navigation + Manipulation Instructions



(a) "go to the blue box"

(b) "move towards the green object"

(c) "travel to the orange object"

Fig. 7. Images of labeled trajectories generated by constraint and environment sampling that form the training and test sets for constraint inference evaluation.

A Natural Language Planner Interface for Mobile Manipulators. Howard, Tellex, and Roy. ICRA 2014.

Configuration/Assembling Instructions

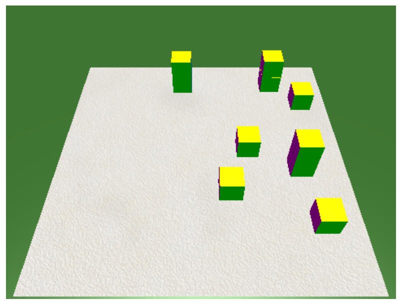
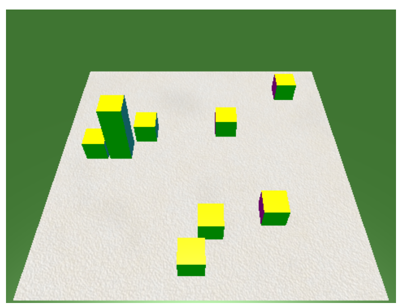
Scene	**Utterance**
	Move the block that is currently located closest to the top left corner to the bottom left of the table, slightly higher than the block in the bottom right corner.
Error:	`7.29 Block lengths`
	Move the block closest to the top left corner so it is above half a block length to the right of the blocks near the lower left corner of the table.
Error:	`0.94 Block lengths`

Table 6: Above are two commands and the worlds they apply to. Below we see the prediction error of our best model.

Natural Language Communication with Robots. Bisk, Yuret, and Marcu. NAACL 2016.

Configuration/Assembling Instructions

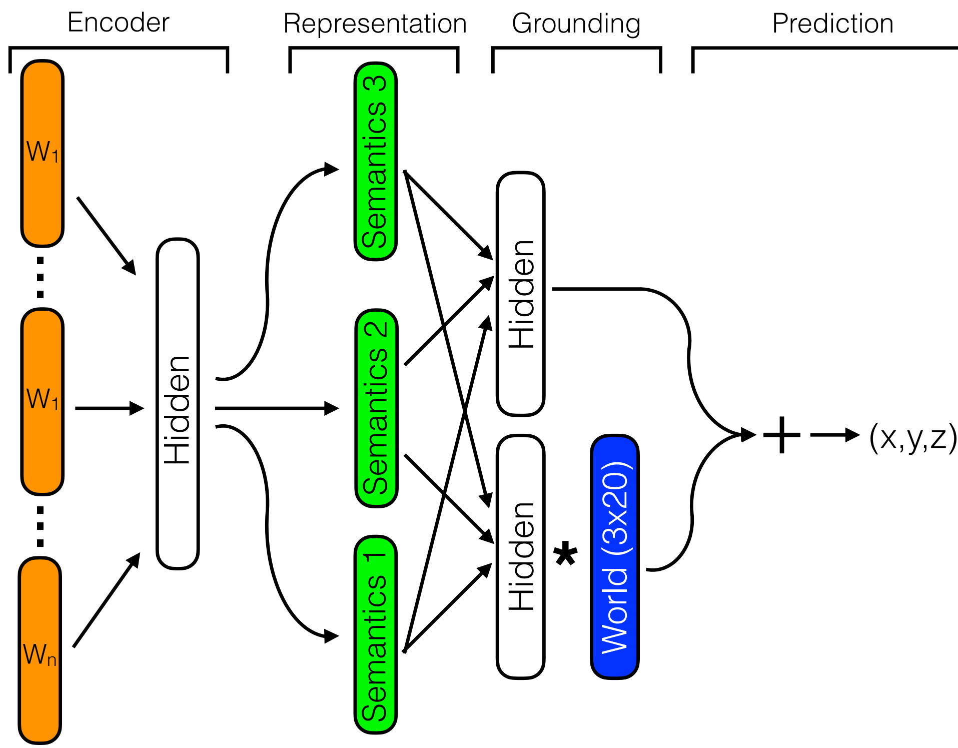


Figure 2: Our models all follow the above architecture. 1-Hot word vectors (orange) are fed as input to a Feed-Forward or Recurrent Neural Network for encoding. A semantic representation is extracted (green), which in conjunction with knowledge of the world (blue) is grounded to predict an action.

Configuration/Assembling Instructions

		MNIST Patterns with labeled blocks							Random Patterns with blank blocks						
		Source		Target					Source		Target				
		Med	Mean	Med	Mean	S	R	D	Med	Mean	Med	Mean	S	R	D
	Human Performance	0.00	0.00	0.21	0.53	100			0.00	0.30	0.37	1.39	93		
	Oracle	–	–	0.00	0.45	100	100	100	–	–	1.00	1.09	100	100	100
FFN	Discrete Predictions	**0.00**	0.49	**1.09**	2.17	93	69	63	5.28	5.09	5.51	5.46	9	15	32
FFN	Continuous Predictions	0.49	1.00	1.59	2.42				4.25	4.04	3.86	**3.93**			
FFN	End-to-End	0.02	**0.38**	1.14	**1.81**				**3.45**	**3.52**	**3.60**	3.94			
RNN	Discrete Predictions	**0.00**	**0.14**	**0.00**	**0.98**	98	92	78	5.29	5.00	5.51	5.57	10	7	46
RNN	Continuous Predictions	0.47	0.64	1.23	1.60				4.16	4.05	3.71	3.87			
RNN	End-to-End	0.03	0.19	0.53	1.05				**3.29**	**3.47**	**3.60**	**3.70**			
	Center Baseline	–	–	3.46	3.43	100			–	–	4.09	4.06	100		
	Random Baseline	6.37	6.49	6.12	6.21	5	5	11	4.90	4.97	5.51	5.44	10	11	12

Table 4: Model error when trained on only the subset of the data with decorated blocks or blank blocks. Where appropriate S, R, and D are the model's predictive accuracy at identifying the Source, Reference and Direction. All models are evaluated on the Median and Mean prediction error the source block and its final target location. Distances are presented in block-lengths.

Natural Language Communication with Robots. Bisk, Yuret, and Marcu. NAACL 2016.

Configuration/Assembling Instructions

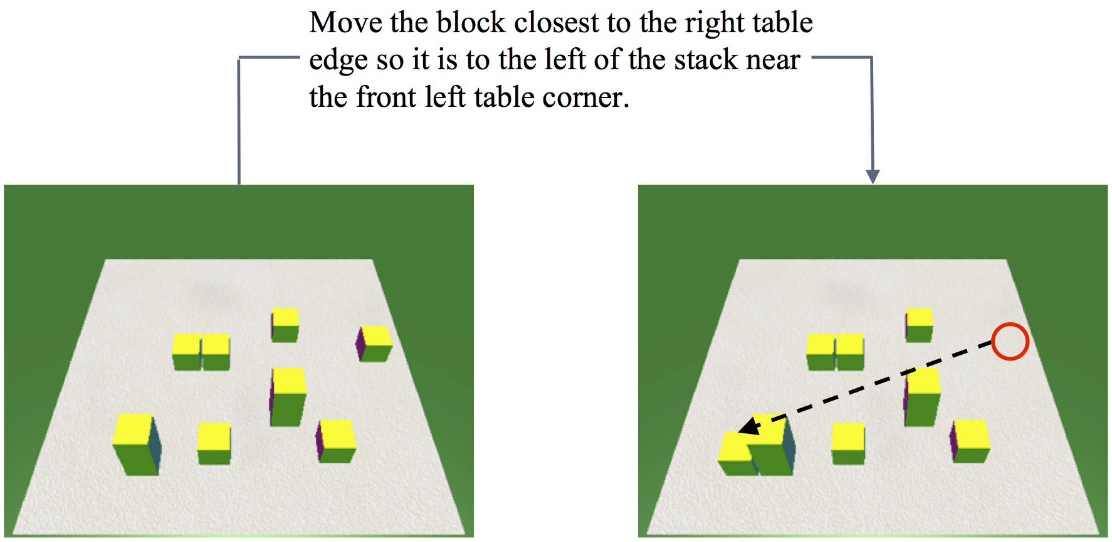


Figure 1: An example of the configuration instruction understanding task (based on blank-labeled blocks). Our model is able to correctly predict the source block and the target position in this case.

Configuration/Assembling Instructions

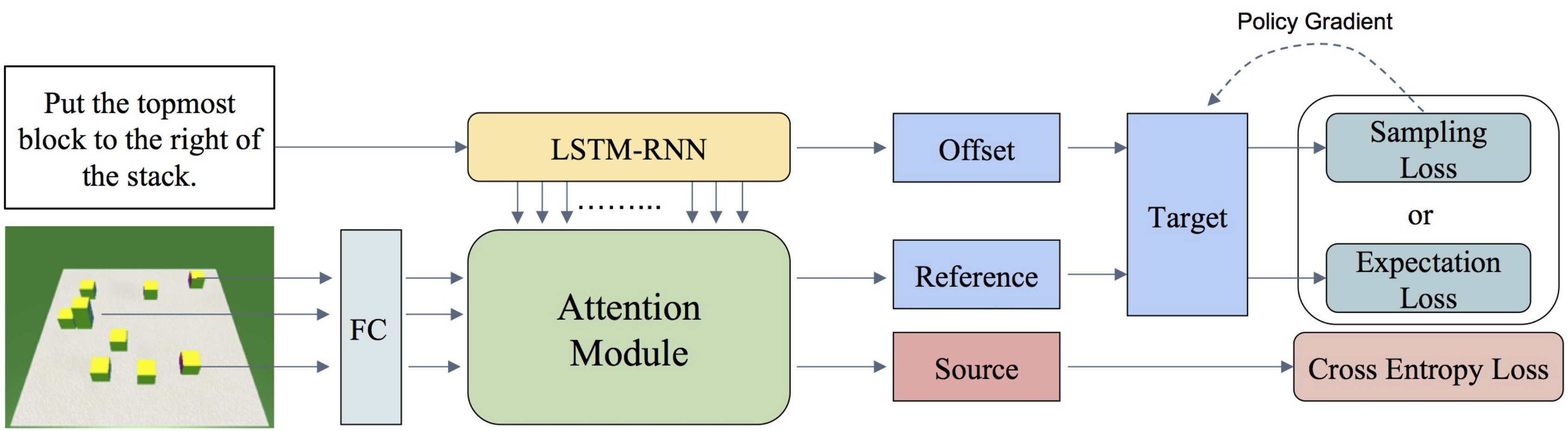


Figure 2: Our overall model for the assembly instruction understanding task, showing instruction and world representation learning, language-to-block alignment modules, and source and target (expectation vs. sampling) loss functions.

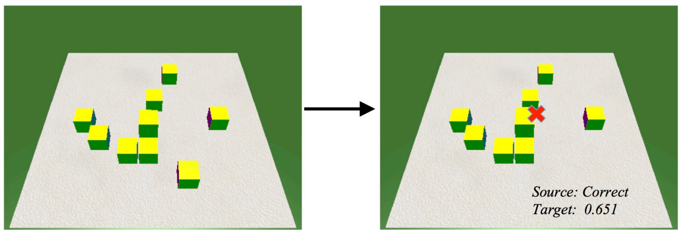
Configuration/Assembling Instructions

Model	SOURCE			TARGET	
	Accuracy	Median	Mean	Median	Mean
End-to-End FFN (Bisk, Yuret, and Marcu 2016)	9.0%	3.45	3.52	3.60	3.94
End-to-End RNN (Bisk, Yuret, and Marcu 2016)	10.0%	3.29	3.47	3.60	3.70
Our Expectation Model	56.1%	0.00	2.21	2.78	3.07
Our Sampling Model	56.3%	0.00	2.18	3.12	3.18
Our Expectation Model w/ Ensemble	56.6%	0.00	2.12	2.65	2.91
Our Sampling Model w/ Ensemble	56.8%	0.00	2.11	2.71	2.90

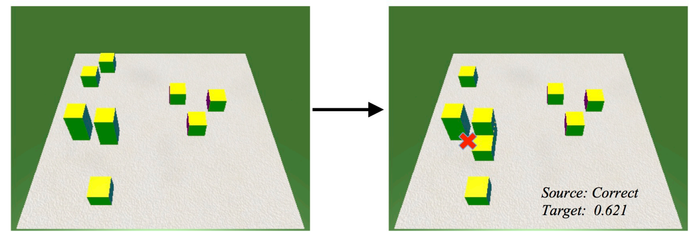
Source-Target Inference Models for Spatial Instruction Understanding. Tan and Bansal. AAAI 2018.

Configuration/Assembling Instructions

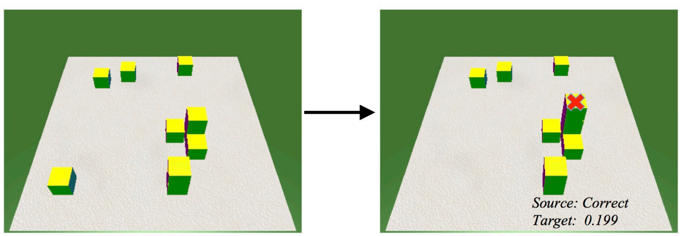
The box in the bottom right, slightly right of center, moves one space north of the tower.



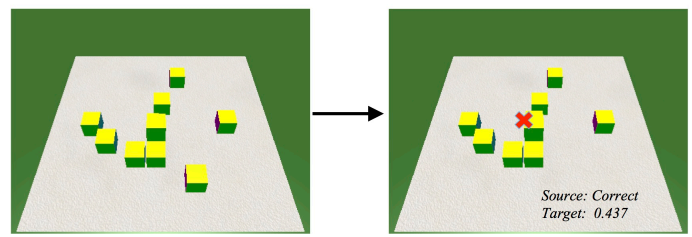
Move the highest block down to below and in front of the right stack of blocks.



Take the leftmost front block and place it on top of the stack of two blocks furthest to the back.



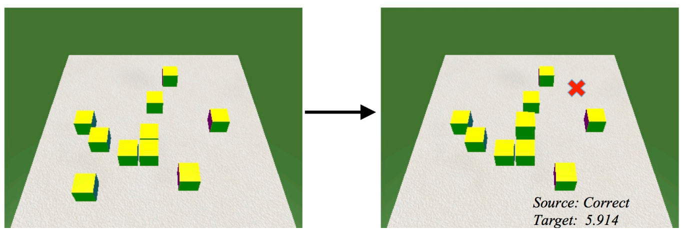
Take the block from the last row and hide it behind the tower.



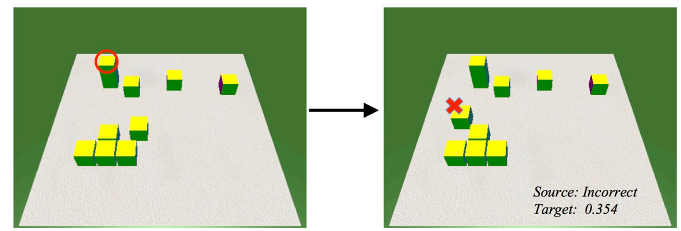
Positive Examples

Configuration/Assembling Instructions

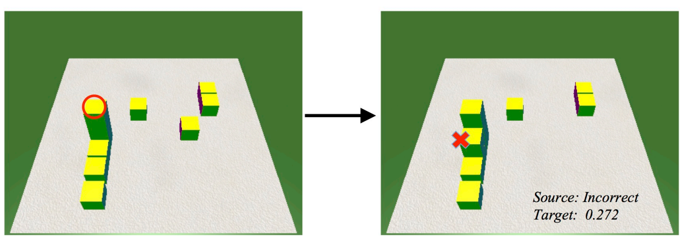
Move the block closest to the bottom left corner so that it is on top of the block at the top of the backwards L .



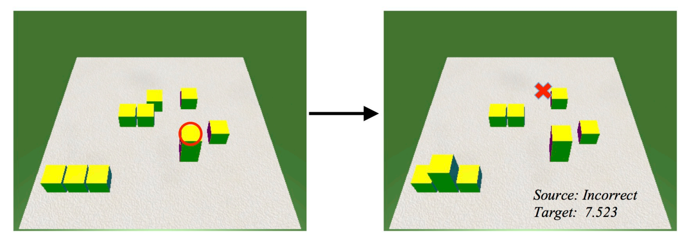
The box next to the Tetris structure moves two spaces left and one half up.



Slide the block left of the two in the top right over and on top of the block in front of the tower.



In the 3-piece-long line, the middle box takes a second story from the middle box in the top row.



Negative Examples

Recipe Instruction Following

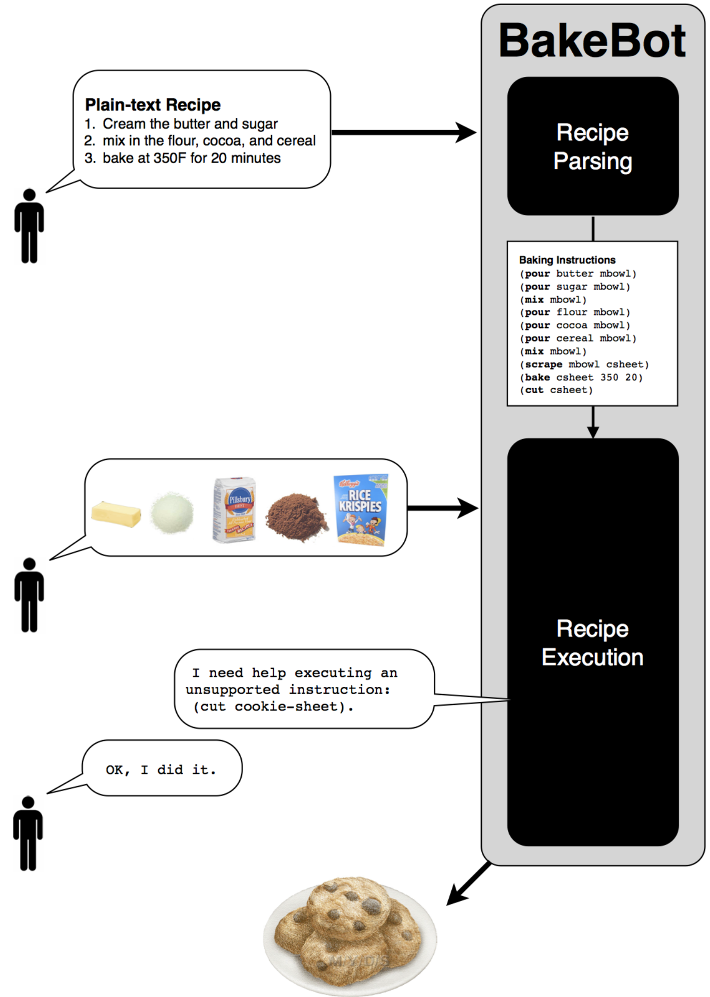


Fig. 2 The human interaction with the BakeBot system for recipe execution. First the person provides the plain-text recipe and the measured ingredients. Then BakeBot infers a sequence of baking primitives to execute that correspond to following the recipe. If BakeBot encounters an unsupported baking primitive, it asks its human partner for help executing the instruction. The end result is baked cookies.

Interpreting and Executing Recipes with a Cooking Robot. Bollini, Tellex, Thompson, Roy, Rus. ISER 2012.

Recipe Instruction Following

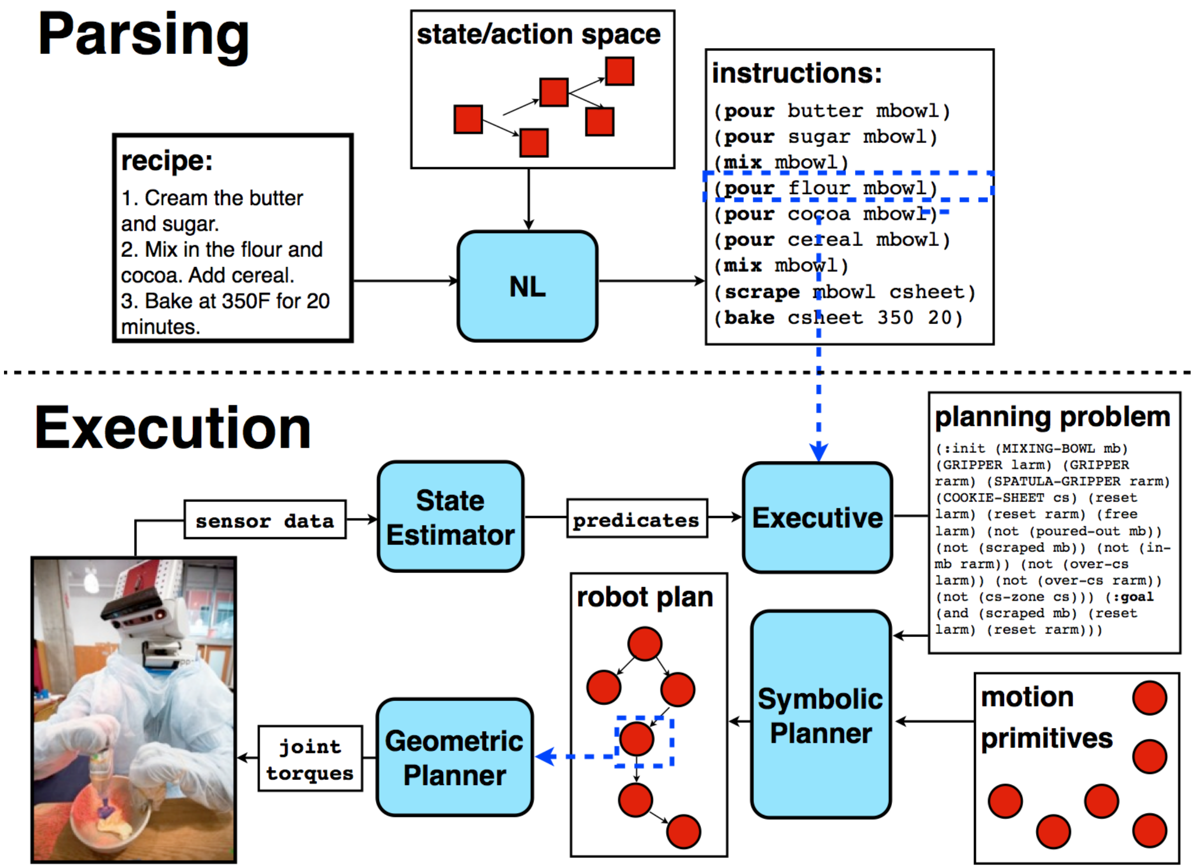


Fig. 3 Architecture of the BakeBot system. The NL system processes the plain text recipe, producing a high-level plan which is sent to the robot. For each instruction in the high-level plan, the motion planner assembles a motion plan and executes it on the PR2 robot.

http://projects.csail.mit.edu/video/research/robo/bakebot_final.mp4

Interpreting and Executing Recipes with a Cooking Robot. Bollini, Tellex, Thompson, Roy, Rus. ISER 2012.

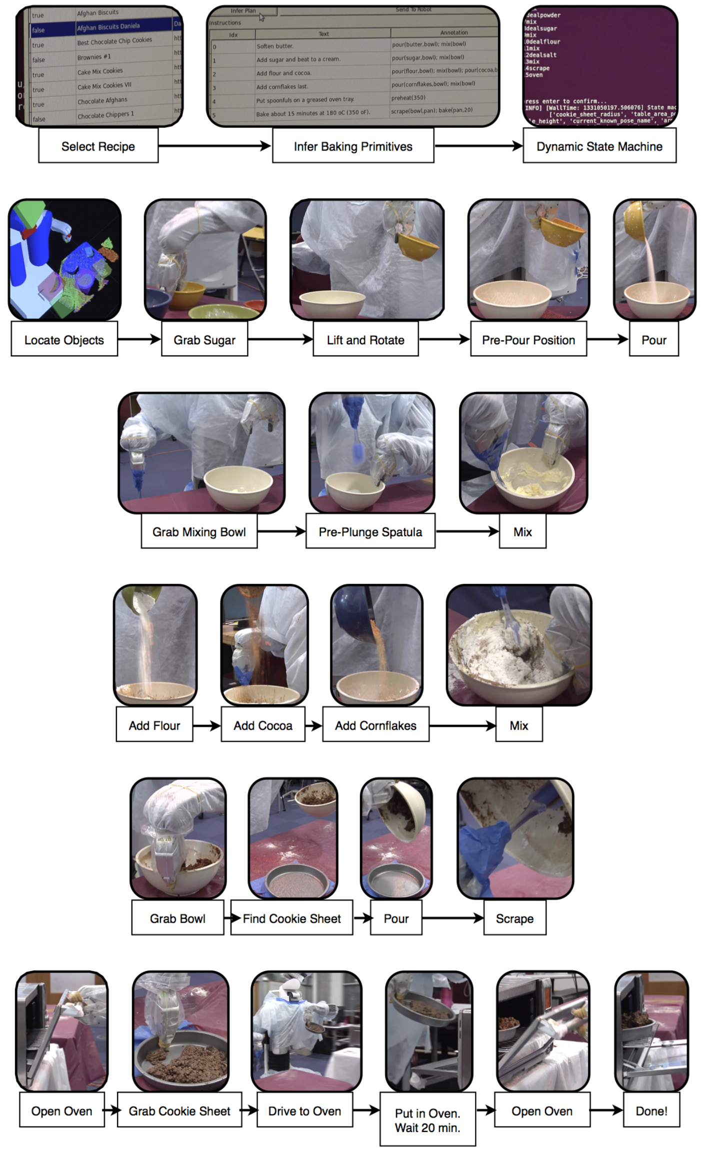
Recipe Instruction Following

Recipe Text	**Inferred Action Sequence**
Afghan Biscuits	
200g (7 oz) butter	
75g (3 oz) sugar	
175g (6 oz) flour	
25g (1 oz) cocoa powder	
50g cornflakes (or crushed weetbix)	
Soften butter.	$pour(butter, bowl); mix(bowl)$
Add sugar and beat to a cream.	$pour(sugar, bowl); mix(bowl)$
Add flour and cocoa.	$pour(flour, bowl); pour(cocoa, bowl)$
Add cornflakes last.	$pour(cornflakes, bowl); mix(bowl)$
Put spoonfuls on a greased oven tray.	$scrape()$
Bake about 15 minutes at 180°C (350°F).	$preheat(350); bake(pan, 20)$

Fig. 4 Text from a recipe in our dataset, paired with the inferred action sequence for the robot.

Interpreting and Executing Recipes with a Cooking Robot. Bollini, Tellex, Thompson, Roy, Rus. ISER 2012.

Recipe Instruction Following



Interpreting and Executing Recipes with a Cooking Robot. Bollini, Tellex, Thompson, Roy, Rus. ISER 2012.

Recipe Instruction Following

- Recipes: Tell Me Dave (http://tellmedave.cs.cornell.edu/)

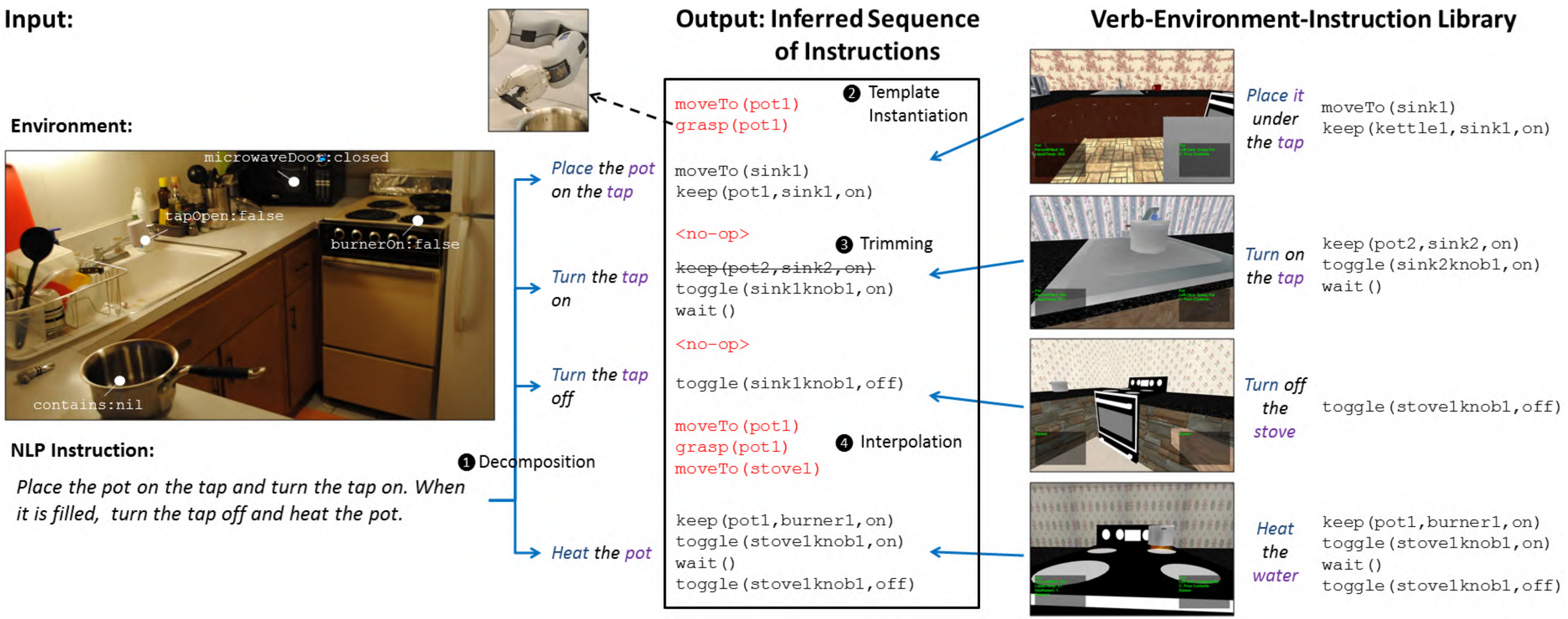


Fig. 1. **Natural Language Instructions to sequence of instructions** for a given new environment. Our approach takes description in natural language and sequences together robotic instructions that are appropriate for a given environment and task. Note that the NL instructions are often ambiguous, and are incomplete, and need to be grounded into the environment.

Tell Me Dave: Context-Sensitive Grounding of Natural Language to Mobile Manipulation Instructions, Misra, Sung, Lee, and Saxena. RSS 2014.

Recipe Instruction Following

- Recipes: Tell Me Dave (http://tellmedave.cs.cornell.edu/)

Fig. 4. **Robot Experiment.** Given the language instruction for making the dessert 'Affogato': *'Take some coffee in a cup. Add icecream of your choice. Finally, add raspberry syrup to the mixture.'*, our algorithm outputs a sequence that the PR2 executes to make the dessert. (Please see the video.)

Tell Me Dave: Context-Sensitive Grounding of Natural Language to Mobile Manipulation Instructions, Misra, Sung, Lee, and Saxena. RSS 2014.

Recipe Instruction Following

- Recipes: RoboBarista (http://robobarista.cs.cornell.edu/)

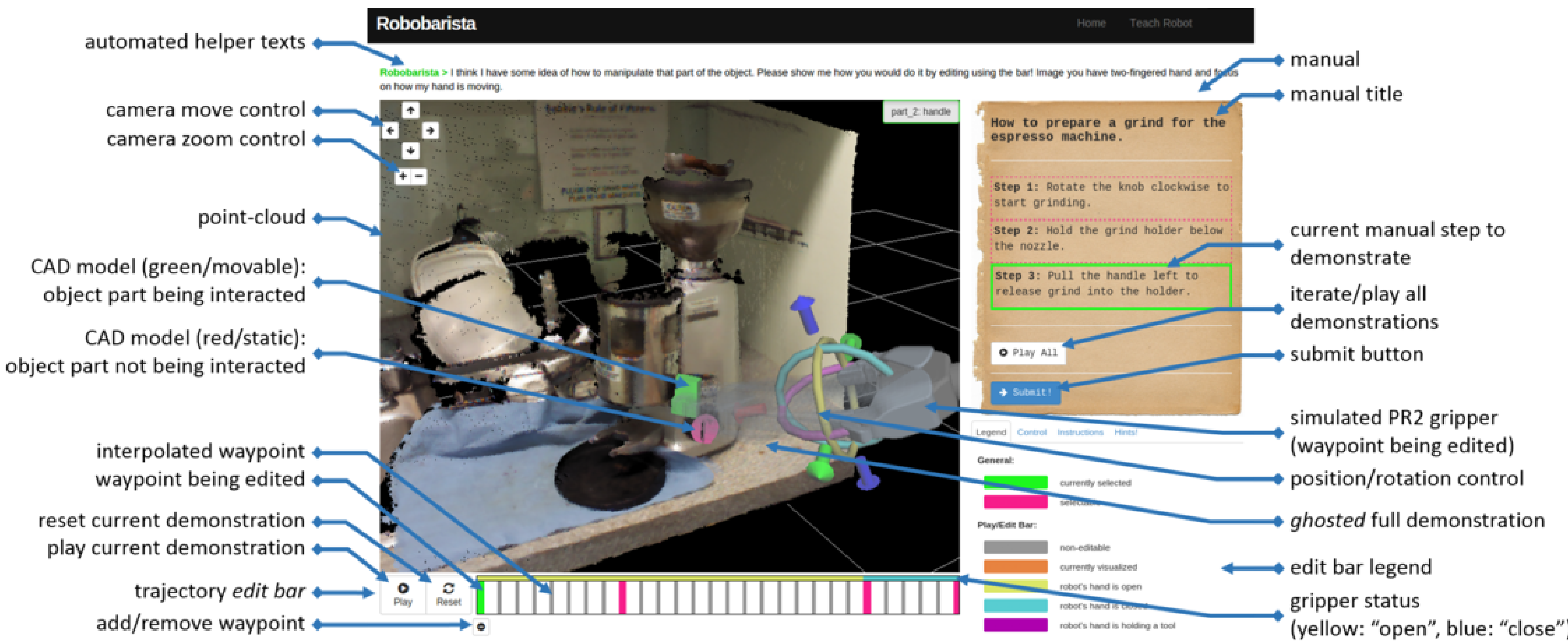


Fig. 5. **Screen-shot of Robobarista,** the crowd-sourcing platform running on Chrome browser. We have built Robobarista platform for collecting a large number of crowd demonstrations for teaching the robot.

Robobarista: Object Part based Transfer of Manipulation Trajectories from Crowd-sourcing in 3D Pointclouds. Sung, Jin, and Saxena. ISRR 2015.

Recipe Instruction Following

- RoboBarista: http://robobarista.cs.cornell.edu/

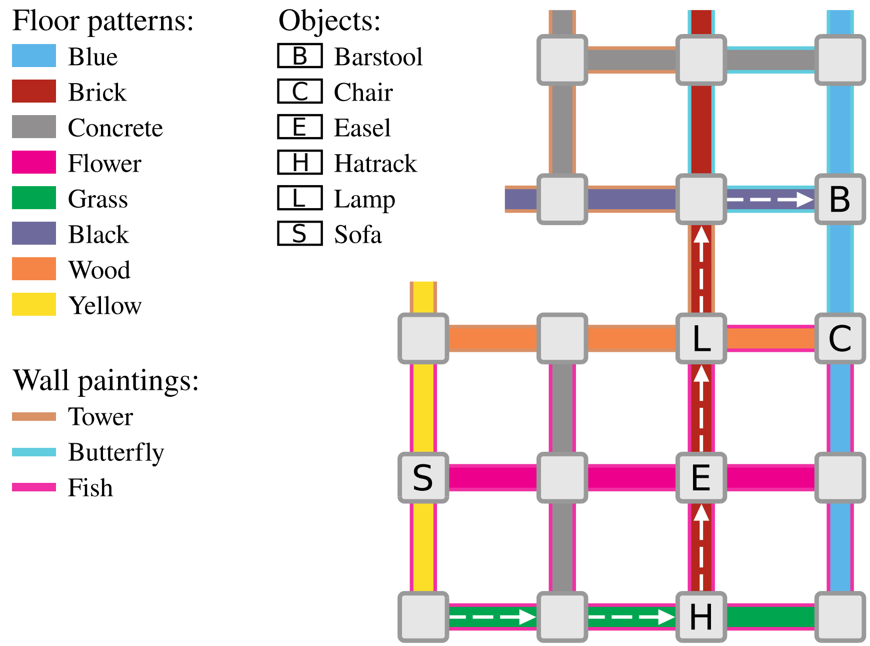
Fig. 10. **Examples of transferred trajectories** being executed on PR2. On the left, PR2 is able to rotate the 'knob' to turn the lamp on. In the third snapshot, using two transferred trajectories, PR2 is able to hold the cup below the 'nozzle' and press the 'lever' of 'coffee dispenser'. In the last example, PR2 is frothing milk by pulling down on the lever, and is able to prepare a cup of latte with many transferred trajectories.

Robobarista: Object Part based Transfer of Manipulation Trajectories from Crowd-sourcing in 3D Pointclouds. Sung, Jin, and Saxena. ISRR 2015.

Navigation Instruction Generation

Fig. 4. Participants' field of view in the virtual world used for the human navigation experiments.

Input: map and path



Output: route instruction

"turn to face the grass hallway. walk forward twice. face the easel. move until you see black floor to your right. face the stool. move to the stool"

Fig. 1. An example route instruction that our framework generates for the shown map and path.

Navigational Instruction Generation as Inverse Reinforcement Learning with Neural Machine Translation. Daniele, Bansal, and Walter. HRI 2017.

Navigation Instruction Generation

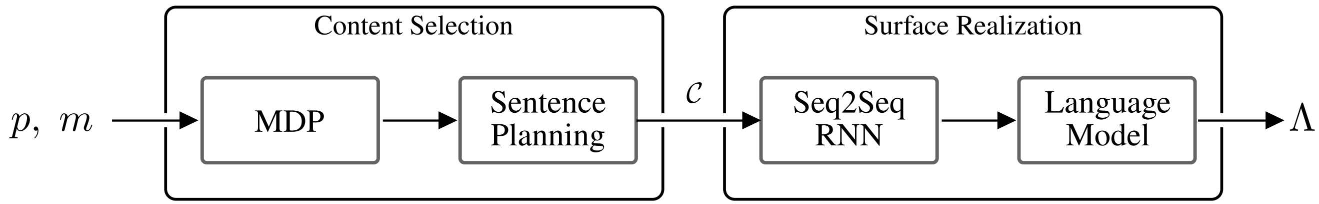


Fig. 2. Our method generates natural language instructions for a given map and path.

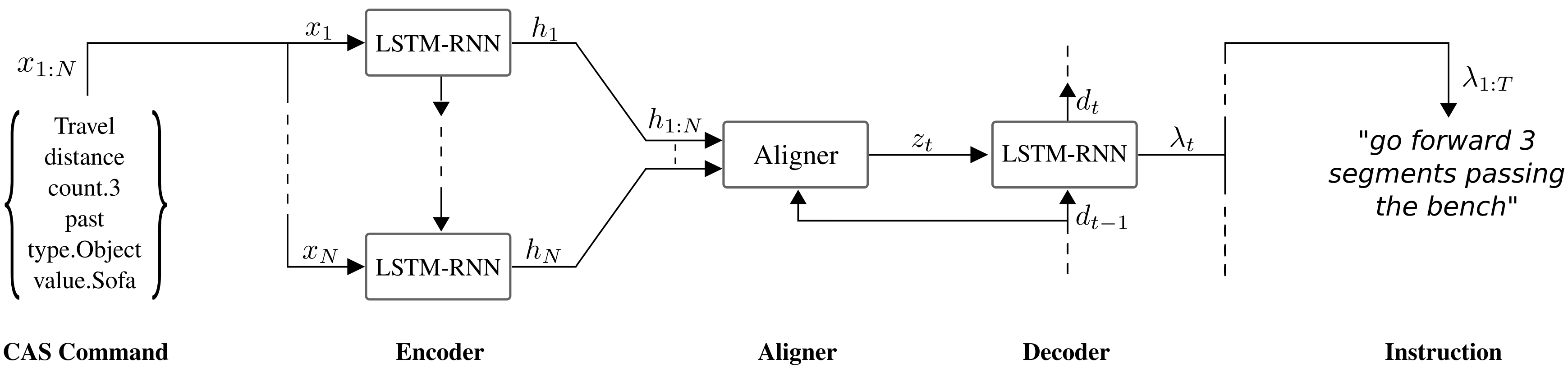


Fig. 3. Our encoder-aligner-decoder model for surface realization.

Navigational Instruction Generation as Inverse Reinforcement Learning with Neural Machine Translation. Daniele, Bansal, and Walter. HRI 2017.

Navigation Instruction Generation

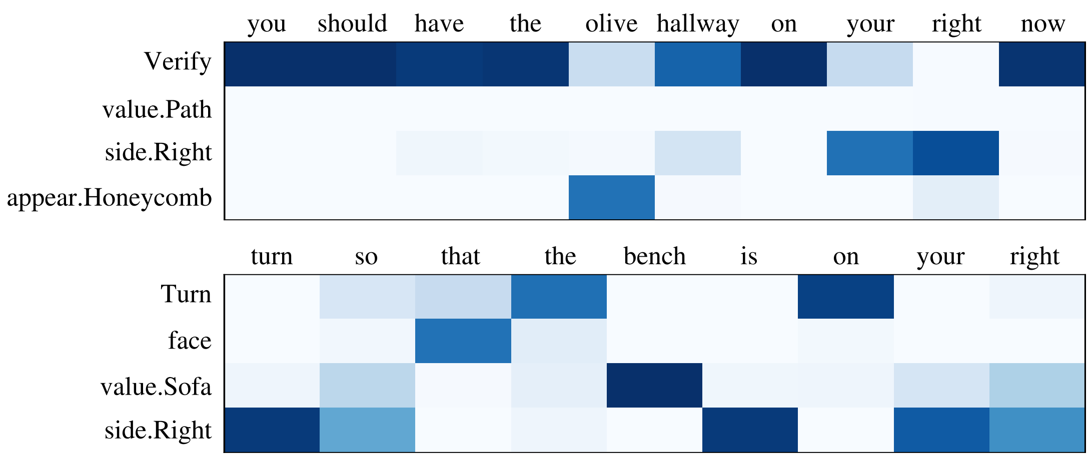
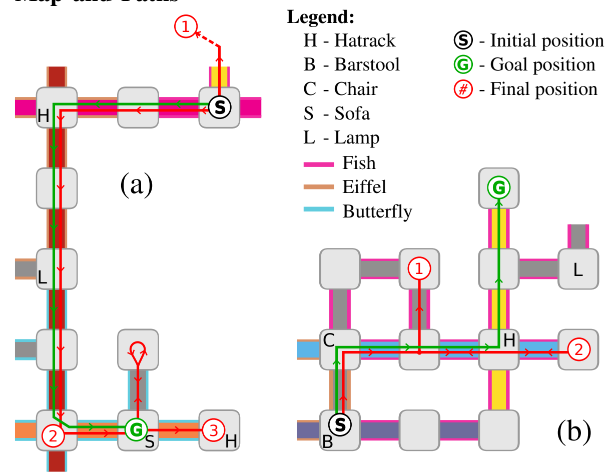


Fig. 5. Alignment visualization for two pairs of CAS (left) and natural language instructions (top). Darker colors denote greater attention weights.

Map and Paths



Instructions

(a)	Human	"with your back to the wall turn left. walk along the flowers to the hatrack. turn left. walk along the brick two alleys past the lamp. turn left. move along the wooden floor to the chair. in the next block is a hatrack"
	Ours	"you should have the olive hallway on your right now. walk forward twice. turn left. move until you see wooden floor to your left. face the bench. move to the bench"
(b)	Human	"head toward the blue floored hallway. make a right on it. go down till you see the fish walled areas. make a left in the fish walled hallway and go to the very end"
	Ours	"turn to face the white hallway. walk forward once. turn right. walk forward twice. turn left. move to the wall"

Fig. 8. Examples of paths from the SAIL corpus that ten participants (five for each map) followed according to instructions generated by humans and by our method. Paths in red are those traversed according to human-generated instructions, while paths in green were executed according to our instructions. Circles with an "S" and "G" denote the start and goal locations, respectively.

Navigational Instruction Generation as Inverse Reinforcement Learning with Neural Machine Translation. Daniele, Bansal, and Walter. HRI 2017.

Language Generation/Dialogue by Robots

- Navigation Dialogue

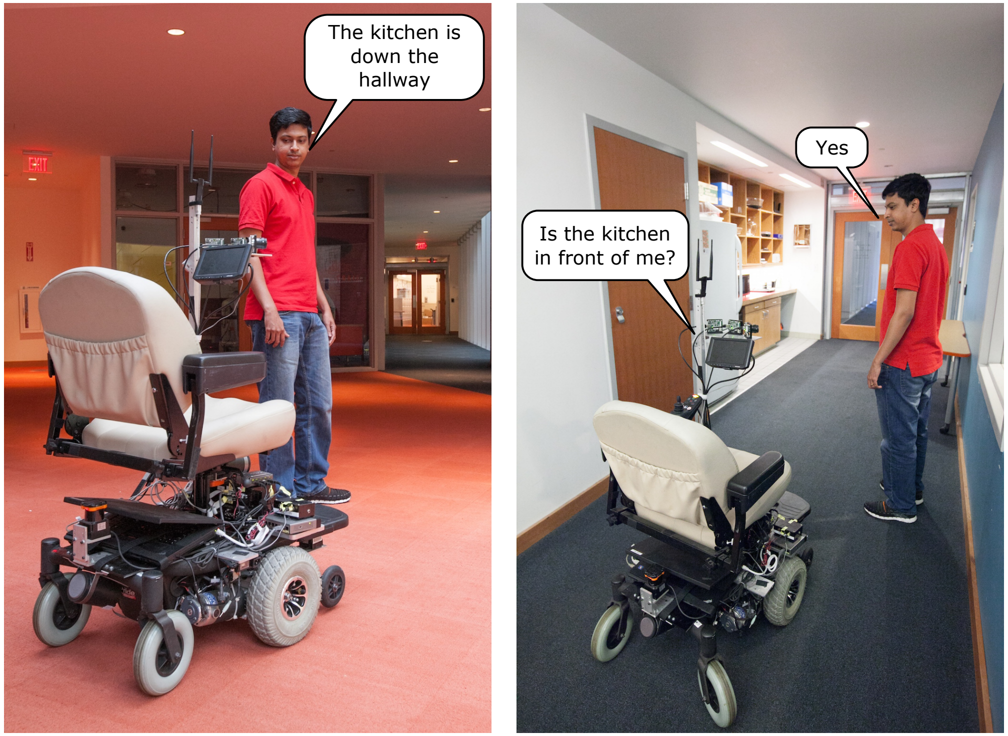
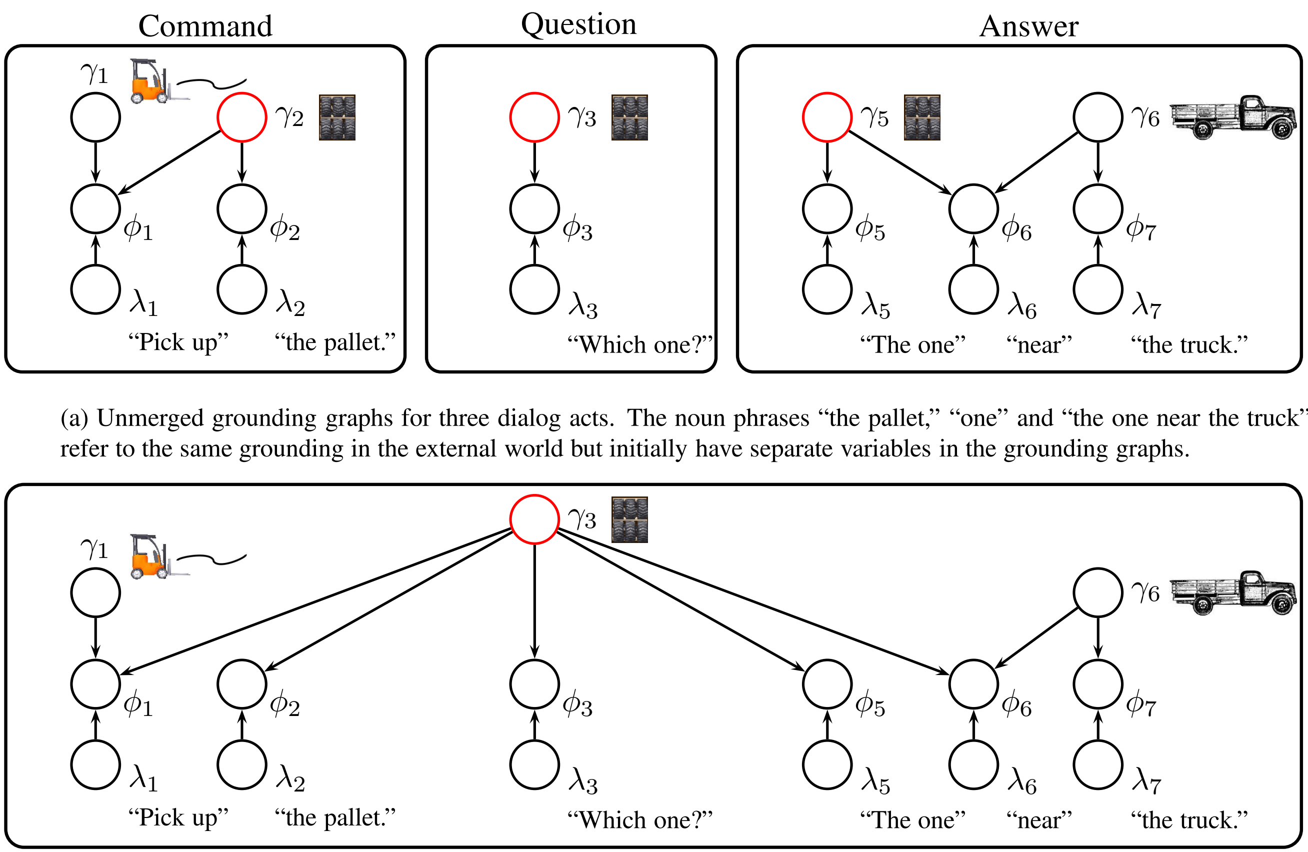


Fig. 1. A user gives a tour to a robotic wheelchair designed to assist residents in a long-term care facility. (Left) The guide provides an ambiguous description of the kitchen's location. (Right) When the robot is near one of the likely locations, it asks the guide a question to resolve the ambiguity.

Information-Theoretic Dialog to Improve Spatial-Semantic Representations. Hemachandra and Walter. IROS 2015.

Language Generation/Dialogue by Robots

- Manipulation Dialogue



(a) Unmerged grounding graphs for three dialog acts. The noun phrases "the pallet," "one" and "the one near the truck" refer to the same grounding in the external world but initially have separate variables in the grounding graphs.

(b) The grounding graph after merging γ_2, γ_3 and γ_5 based on linguistic coreference.

Figure 2. Grounding graphs for a three-turn dialog, before and after merging based on coreference. The robot merges the three shaded variables.

Clarifying Commands with Information-Theoretic Human-Robot Dialog. Deits et al. JHRI 2013.

Language Generation/Dialogue by Robots

- Manipulation Dialogue

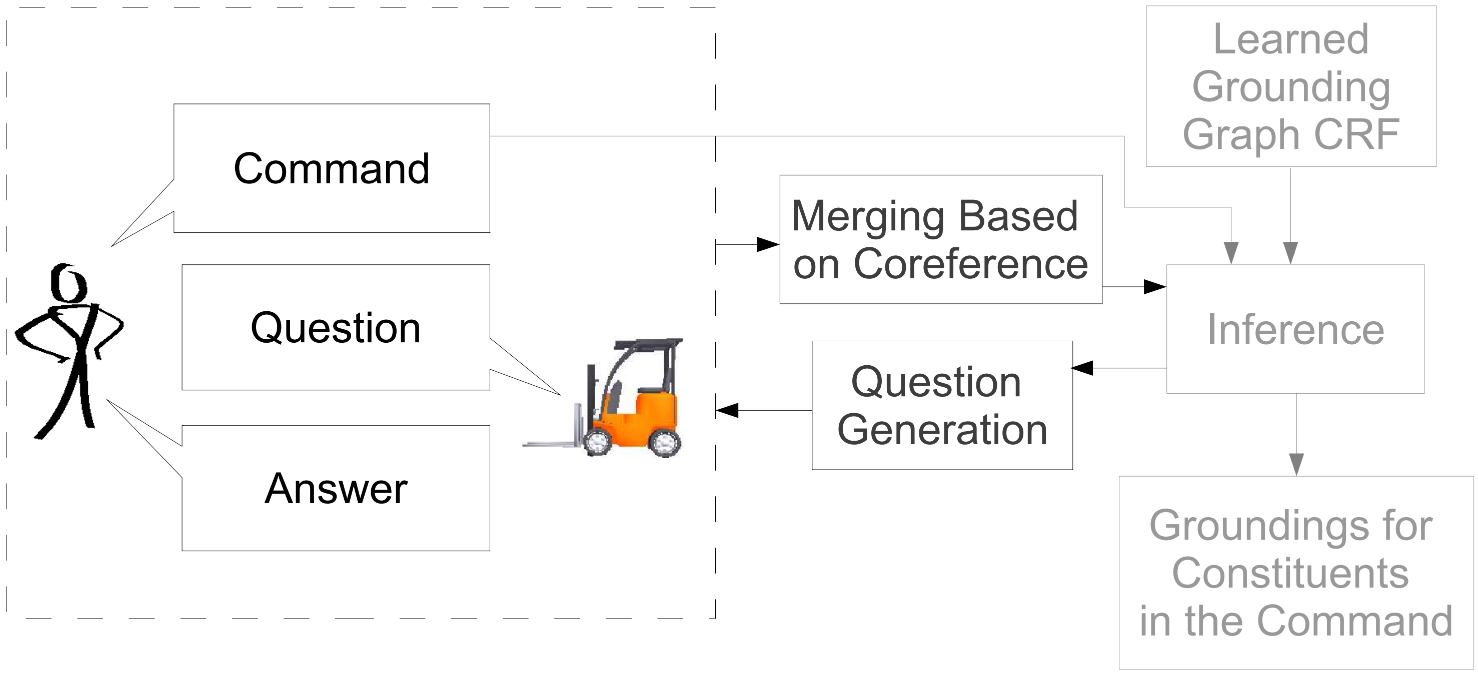


Figure 3. System diagram. Grayed-out blocks show components developed in previous work and are therefore not discussed in detail in this paper; black blocks show the question-asking feedback system new to this paper.

Clarifying Commands with Information-Theoretic Human-Robot Dialog. Deits et al. JHRI 2013.

Language Generation/Dialogue by Robots

- Manipulation Dialogue

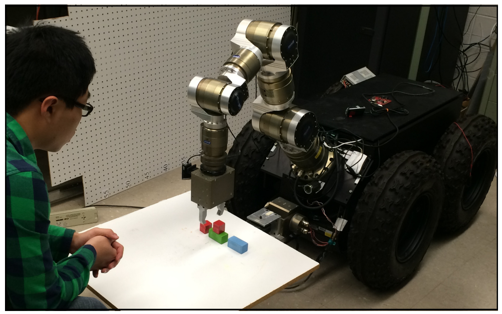
Command:	Move your pallet further right.
Question:	What do the words 'your pallet' refer to?
Answer:	Your pallet refers to the pallet you are currently carrying.
Command:	Move closer to it.
Question:	What does the word 'it' refer to?
Answer:	It refers to the empty truck trailer.
Command:	Take the pallet and place it on the one to the left.
Question:	What do the words 'the one' refer to?
Answer:	The one refers to the empty trailer.
Command:	Place the pallet just to the right of the other pallet.
Question:	What do the words 'the pallet' refer to?
Answer:	The wooden crate that the merchandise sits on top of.

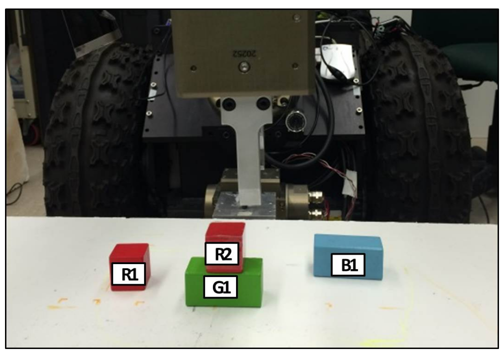
Figure 4. Sample commands, questions, and answers from the corpus.

Clarifying Commands with Information-Theoretic Human-Robot Dialog. Deits et al. JHRI 2013.

Language Generation/Dialogue by Robots

- Learning New Actions via Dialogue





H1: *Stack the blue block on the red block on your right.*
R1: *How to stack, could you teach me?*
H2: *Grab the blue block.*
R2: *What do you mean by grab?*
H3: *Open your gripper.*
R3: *Ok.*
H4: *Move to the blue block.*
R4: *Sure.*
H5: *Close gripper.*
R5: *Alright then.*
H6: *Now you achieve the grab action.*
R6: *Ok, got it.*
H7: *Move the blue block to the top of the right red block.*
R7: *Understood.*
H8: *Open gripper.*
R8: *Sure.*
H9: *Ok, now you are done with the stack action.*
R9: *Ok.*

Figure 1: An example setup and dialogue. Objects are marked with labels only for the illustration purpose.

Back to the Blocks World: Learning New Actions through Situated Human-Robot Dialogue. She, Yang, Cheng, Jia, Chai, Xi. SigDial 2014.

Language Generation/Dialogue by Robots

TASK TO COMPLETE

Dave Daniel wants the item in slot 5.

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

Final Task

DIRECTORY

People:

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

Items available to robot:

Figure 3: The Mechanical Turk interface for the delivery task. This abridged conversation is from a Turker in training batch 0, when the system had access to only the seed lexicon. Because of this conversation, the agent learned that "calender" and "planner" mean "calendar" during retraining.

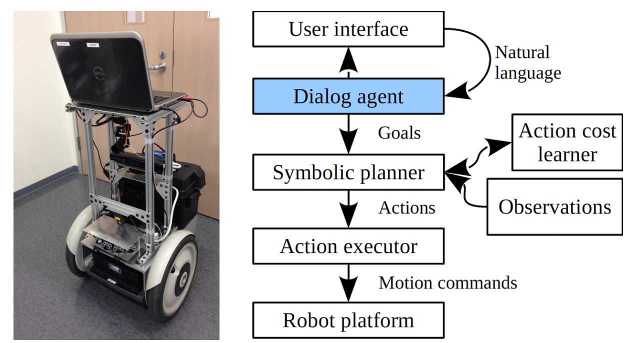


Figure 5: Left: Robot platform (Segbot) used in experiments. **Right:** Segbot architecture, implemented using Robot Operating System (ROS).

Learning to Interpret Natural Language Commands through Human-Robot Dialog. Thomason, Zhang, Mooney and Stone. IJCAI 2015.