Toward Understanding Natural Language Directions

Video

Motivating Example

- **Robot** Someone is on the way to get you out of here. Are there any other people around who need help?
- **Victim** I saw someone in the main lobby.
- **Robot** Where is the main lobby?
- **Person** Exit this room and turn right. Go down the hallway past the elevators. The lobby is straight ahead.
- **Robot** Understood.

Data Corpus

- Data collection
  - 15 visitors wrote 10 sets of directions each (150 total)
  - Each visitor tries to follow someone else's directions to check quality
    - Best direction giver – 100% followable instructions
    - Worst direction giver – 30% followable instructions
  - Only landmarks shown on predetermined map could be used

With your back to the windows, walk straight through the door near the elevators. Continue to walk straight, going through one door until you come to an intersection just past a white board. Turn left, turn right, and enter the second door on your right (sign says “Administrative Assistant”).

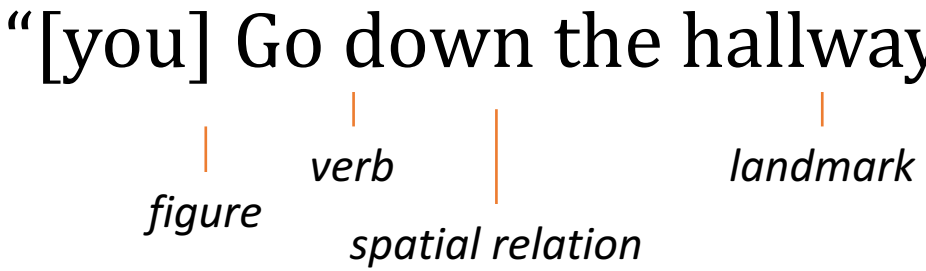
Exploit the Structure of Language

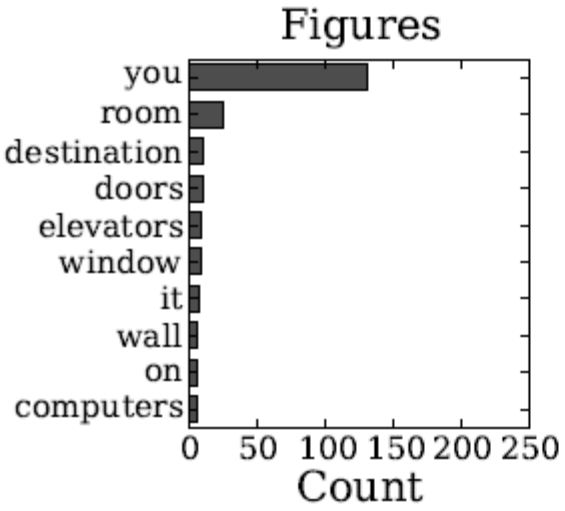
- Directions are:
  - Sequential
  - Contain references to landmarks
  - Contain spatial relations (though, past, etc)
  - Contain verbs

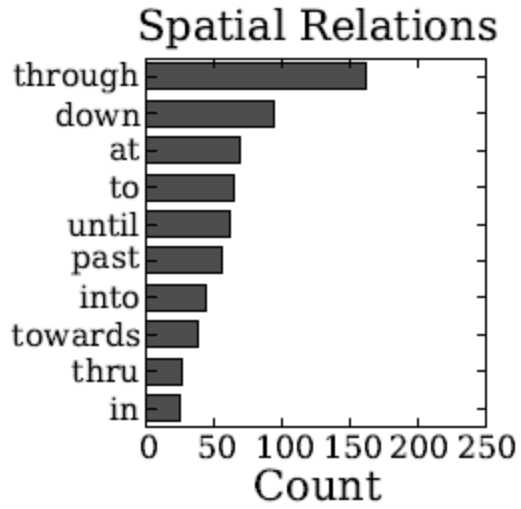
Spatial Description Clause (SDC)

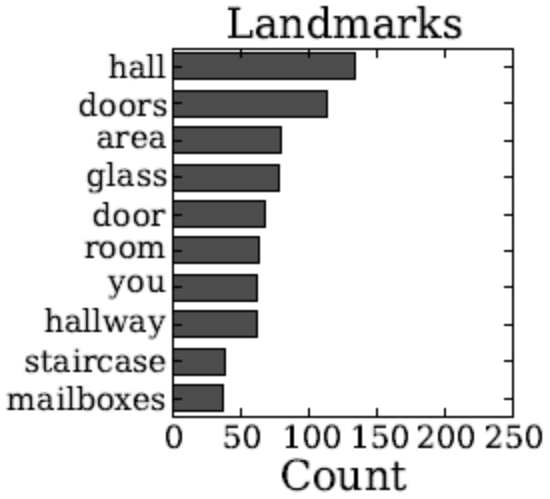
- figure (the subject of the sentence)
- verb (an action to take)
- landmark (an object in the environment)
- spatial relation (a geometric relation between the landmark and the figure)

- Any of these fields can be unlexicalized and therefore only specified implicitly.





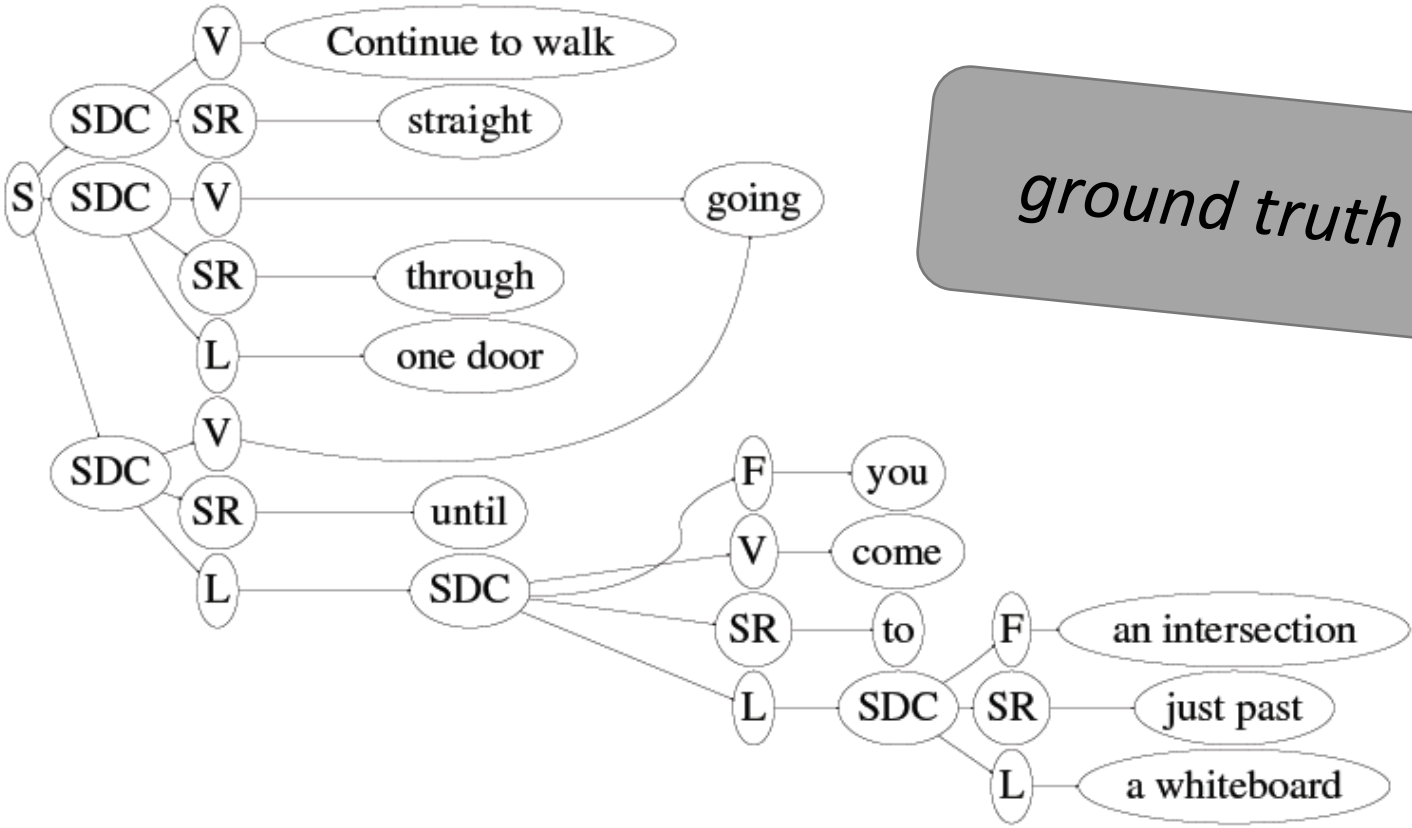




Most frequent words in each SDC field from the corpus if 150 directions (hand annotated).

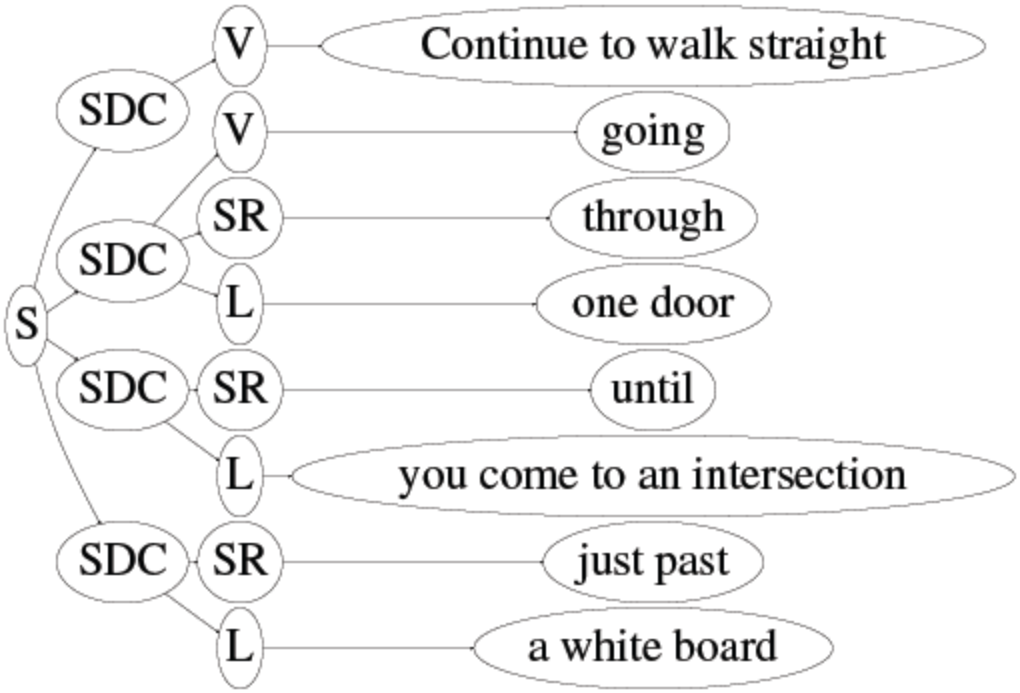
Process

- Automatically extract SDCs from text (CRFs)
- Ground each part in the environment

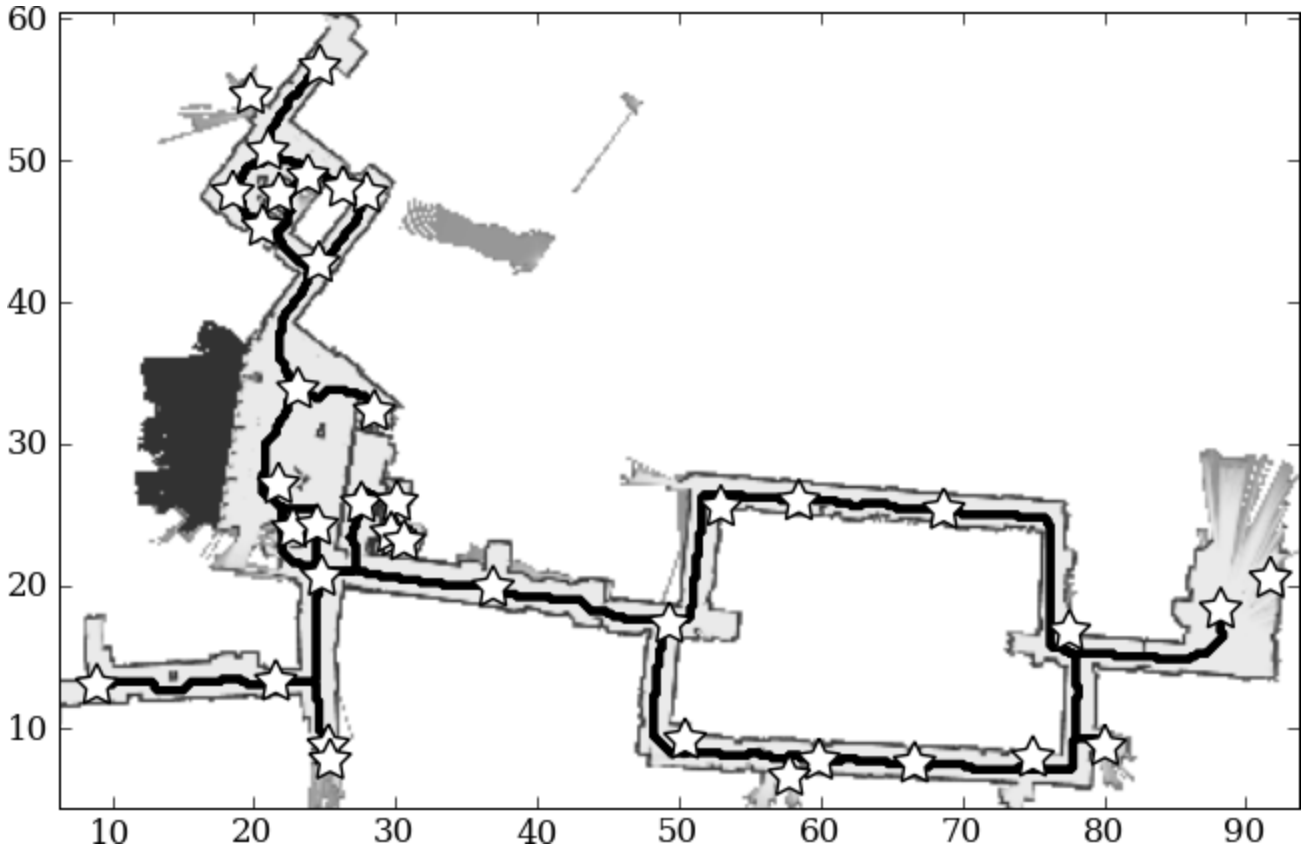


ground truth

automatic


S
SDC
V
Continue to walk straight
SDC
V
going
SR
through
L
one door
SDC
SR
until
L
you come to an intersection
SDC
SR
just past
L
a white board

Topological Map



$$\arg\max_{P} p(P, S|O) = p(S|P, O) \times p(P|O)$$

Conditional independence of three disjoint variables: once O is known, knowing S can no longer influence the probability of P.

They add an additional assumption that the path is independent of the objects. Which leads to:

$$p(P, S|O) \approx p(\mathrm{sdc}_1 \ldots \mathrm{sdc}_M | v_1 \ldots v_{M+1}, O) \times p(v_1 \ldots v_{M+1})$$

$$p(P,S|O) \approx p(\text{sdc}_1 \ldots \text{sdc}_M | v_1 \ldots v_{M+1}, O) \times p(v_1 \ldots v_{M+1})$$

Additional simplifying assumptions (standard Markovian):

1) an SDC depends only on the current transition v_i , v_{i+1}
2) the next viewpoint v_{i+1} depends only on previous viewpoints.

$$p(P,S|O) = \left[\prod_{i=1}^{M} p(\text{sdc}_i | v_i, v_{i+1}, O)\right] \times \left[\prod_{i=1}^{M} p(v_{i+1} | v_i \ldots v_1)\right] \times p(v_1)$$

Obtain the probabilities from their labeled training data:

$$\begin{aligned} p(\text{sdc}_i | v_i, v_{i+1}, O) &= p(f_i, a_i, s_i, l_i, | v_i, v_{i+1}, O) \\ &\approx p(f_i | v_i, v_{i+1}, o_1 \ldots o_K) \times p(a_i | v_i, v_{i+1}) \times \\ &\quad p(s_i | l_i, v_i, v_{i+1}, o_1 \ldots o_K) \times p(l_i | v_i, v_{i+1}, o_1 \ldots o_K) \end{aligned}$$

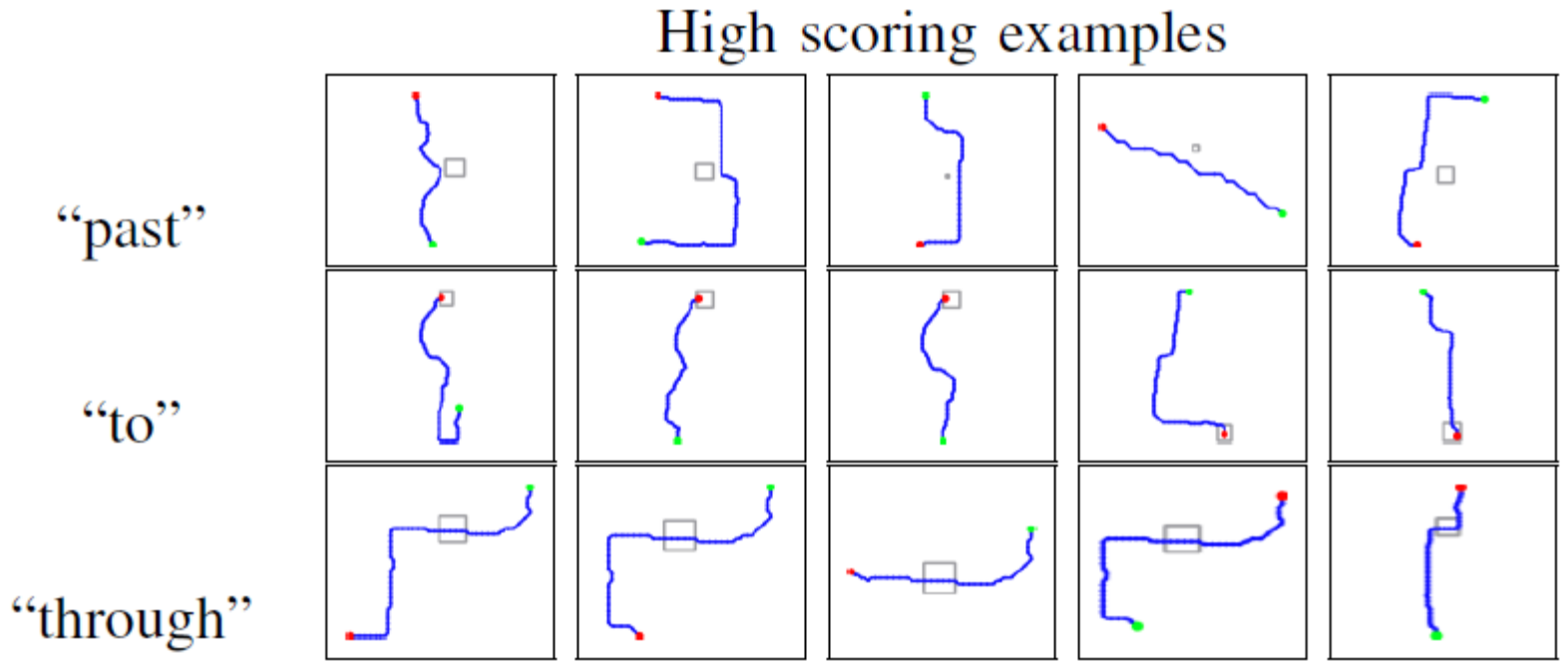
Grounding the figures to physical landmarks

- “the door near the elevator”, “a beautiful view of the domes”

- Download >1M images from Flikr
- Used this dataset to model object co-occurrence
  - $P(kitchen|microwave, toaster)$

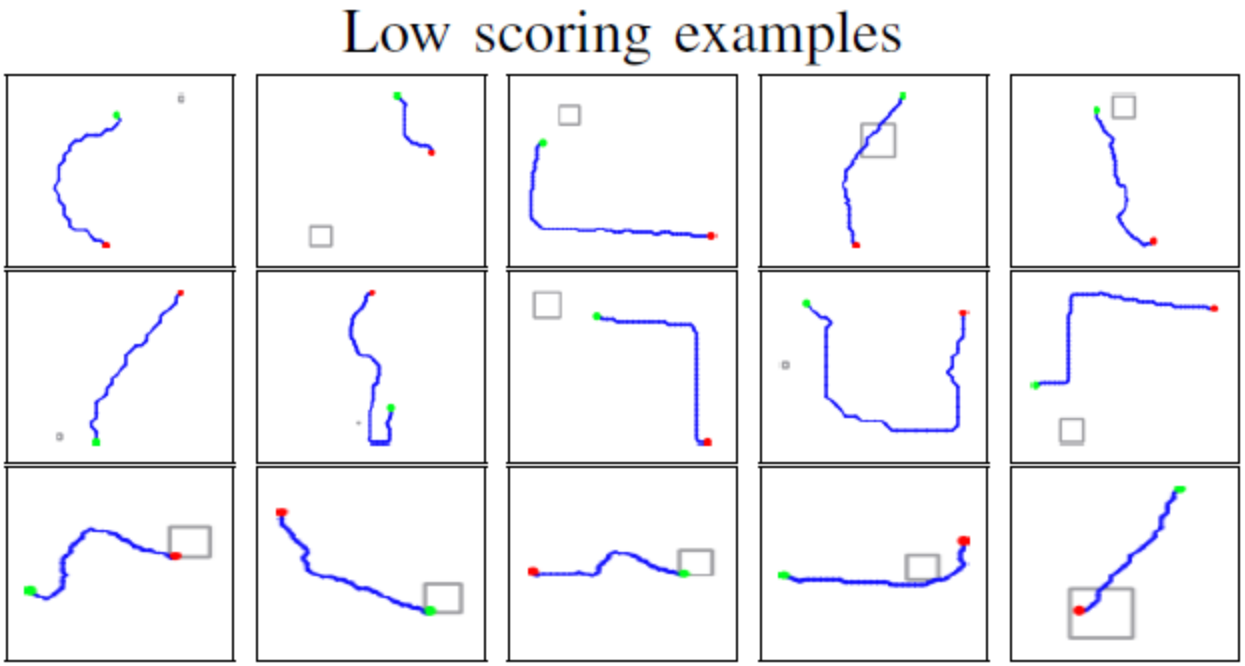
Grounding spatial relations

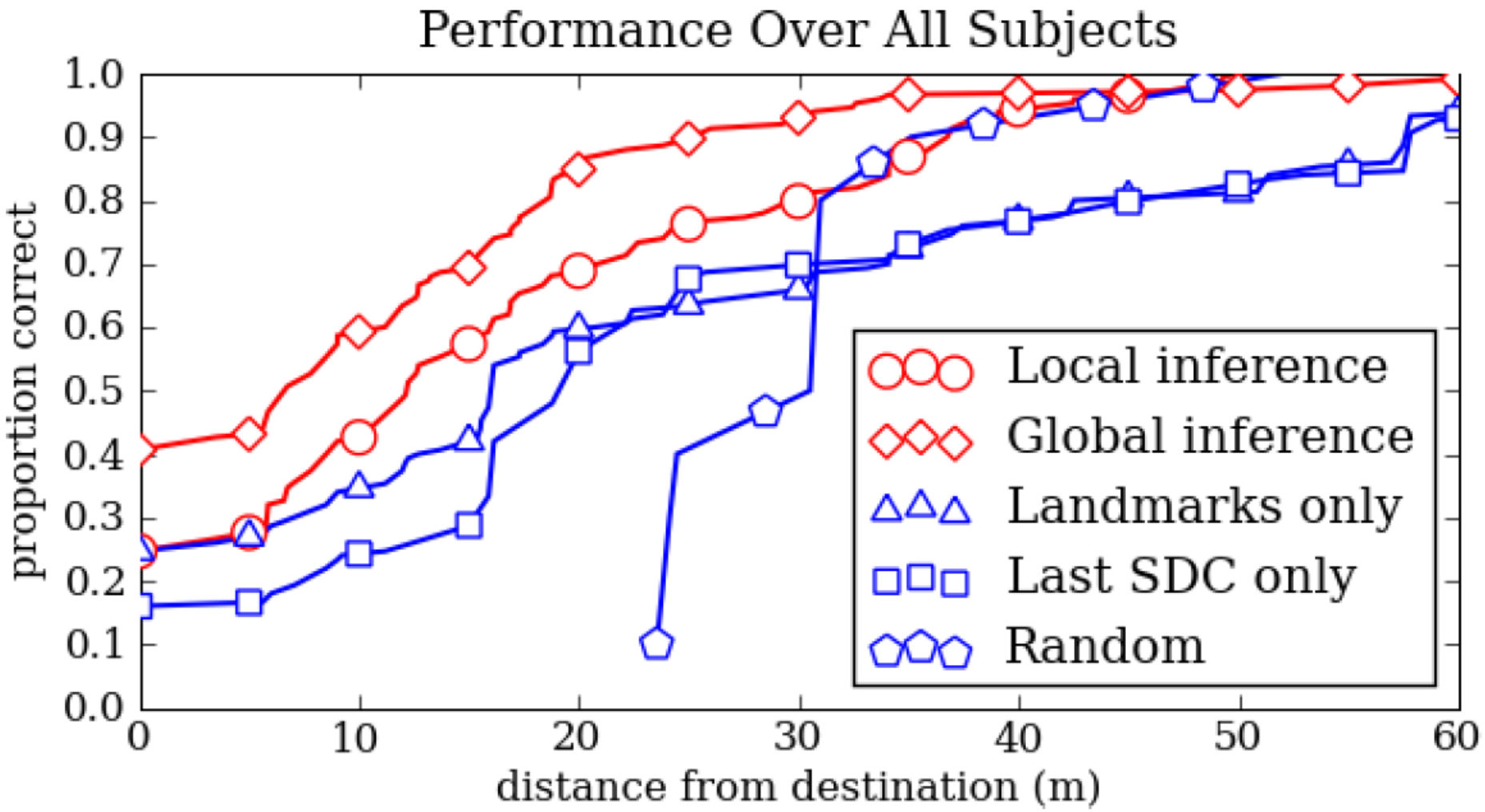
- Hand drawn training examples



Grounding spatial relations

- Hand drawn training examples




Performance Over All Subjects
proportion correct
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
distance from destination (m)
Local inference
Global inference
Landmarks only
Last SDC only
Random

Evaluation

TABLE I
THE PERFORMANCE OF OUR MODELS AT 10 METERS.

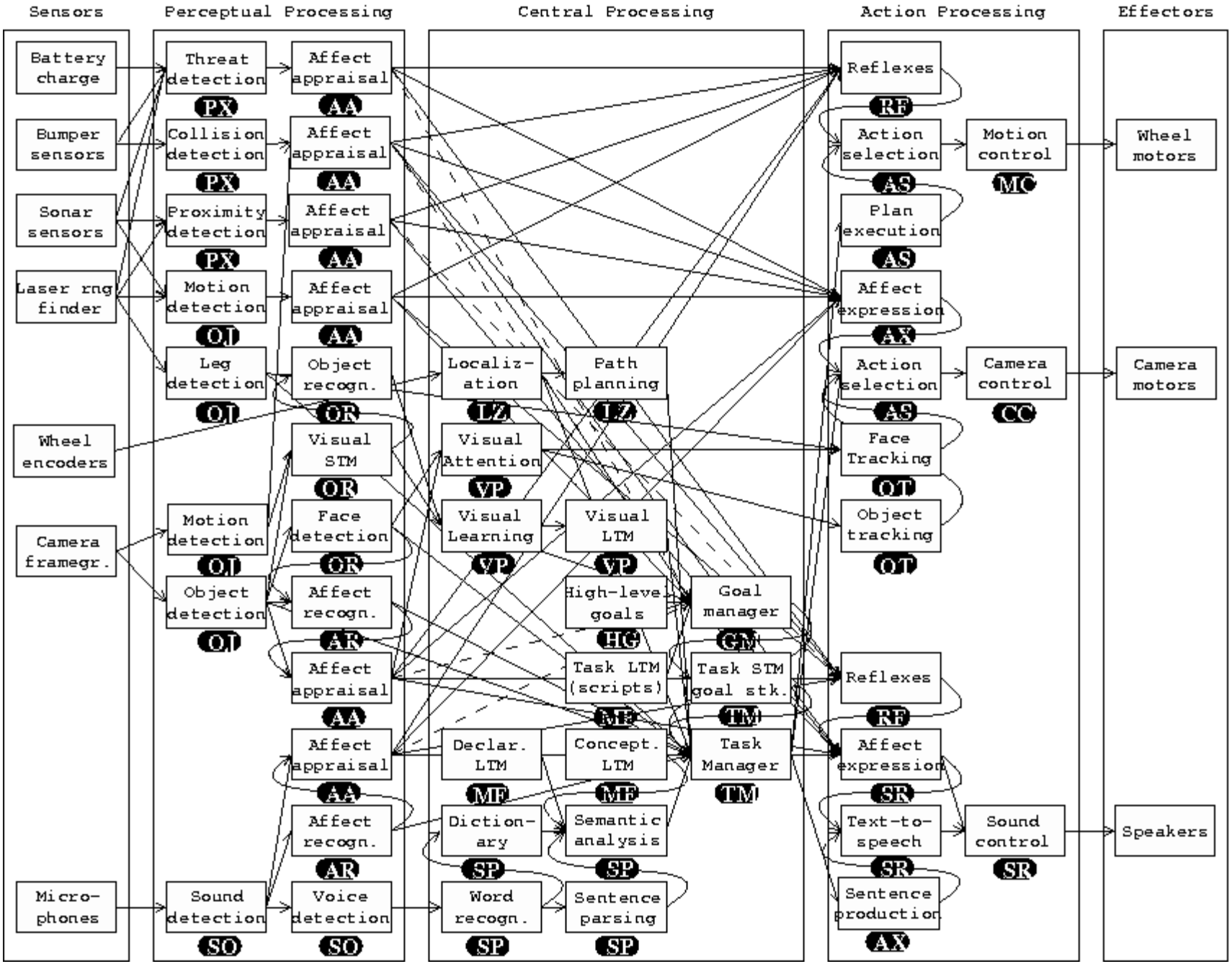
Algorithm	% correct Max Prob	Best Path
Global inference w/spatial relations	48.0%	59.3%
Global inference w/o spatial relations	48.0%	54.7%
Local inference w/ spatial relations	28.0%	42.0%
Local inference w/o spatial relations	26.7%	30.7%
Wei et al. [13]	34.0%	34.0%
Last SDC only	23.0%	24.0%
Random	0.0%	–

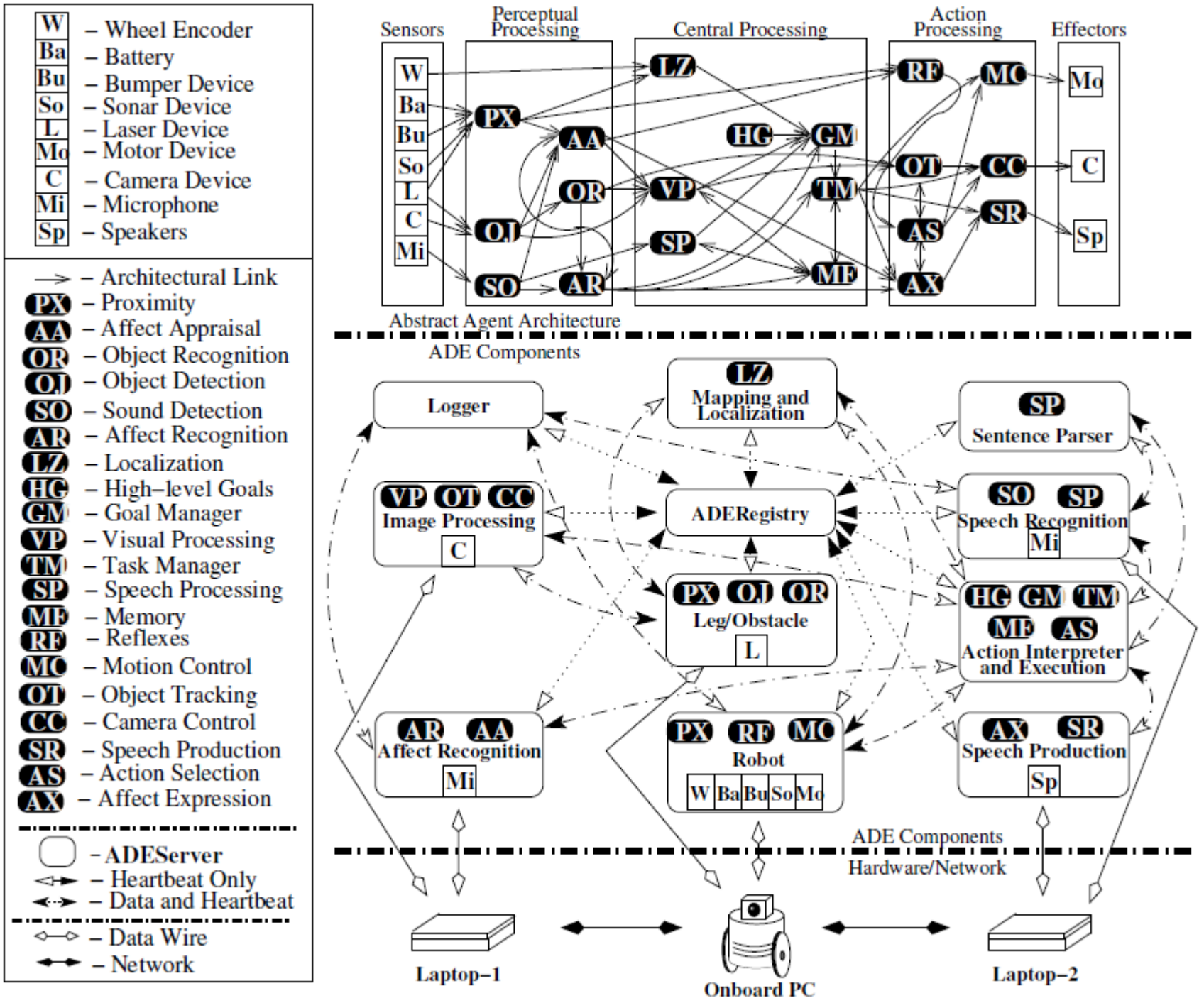
Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation

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

Other related projects

- http://www.youtube.com/user/HRILaboratory?feature=watch


Sensors
Perceptual Processing
Central Processing
Action Processing
Effectors
Battery charge
Bumper sensors
Sonar sensors
Laser rng finder
Wheel encoders
Camera framegr.
Micro-phones
Threat detection
Collision detection
Proximity detection
Motion detection
Leg detection
Motion detection
Object detection
Sound detection
Affect appraisal
Object recogn.
Visual STM
Face detection
Affect recogn.
Voice detection
Localiz-ation
Visual Attention
Visual Learning
Path planning
Visual LTM
High-level goals
Task LTM (scripts)
Declar. LTM
Concept. LTM
Dictionary
Word recogn.
Semantic analysis
Sentence parsing
Goal manager
Task STM goal stk.
Task Manager
Reflexes
Action selection
Plan execution
Affect expression
Face Tracking
Object tracking
Text-to-speech
Sentence production
Motion control
Camera control
Sound control
Wheel motors
Camera motors
Speakers
PX
AA
OJ
OR
VP
LZ
HG
GM
MP
TM
MF
SP
AS
MC
AX
CC
OT
SR
RF
SO
AR


W – Wheel Encoder
Ba – Battery
Bu – Bumper Device
So – Sonar Device
L – Laser Device
Mo – Motor Device
C – Camera Device
Mi – Microphone
Sp – Speakers
– Architectural Link
PX – Proximity
AA – Affect Appraisal
OR – Object Recognition
OJ – Object Detection
SO – Sound Detection
AR – Affect Recognition
LZ – Localization
HG – High–level Goals
GM – Goal Manager
VP – Visual Processing
TM – Task Manager
SP – Speech Processing
ME – Memory
RF – Reflexes
MC – Motion Control
OT – Object Tracking
CC – Camera Control
SR – Speech Production
AS – Action Selection
AX – Affect Expression
– ADEServer
– Heartbeat Only
– Data and Heartbeat
– Data Wire
– Network
Sensors
Perceptual Processing
Central Processing
Action Processing
Effectors
Abstract Agent Architecture
ADE Components
Logger
Mapping and Localization
Sentence Parser
Image Processing
ADERegistry
Speech Recognition
Leg/Obstacle
Action Interpreter and Execution
Affect Recognition
Robot
Speech Production
ADE Components
Hardware/Network
Laptop–1
Onboard PC
Laptop–2