Grounded Language Learning

Ray Mooney

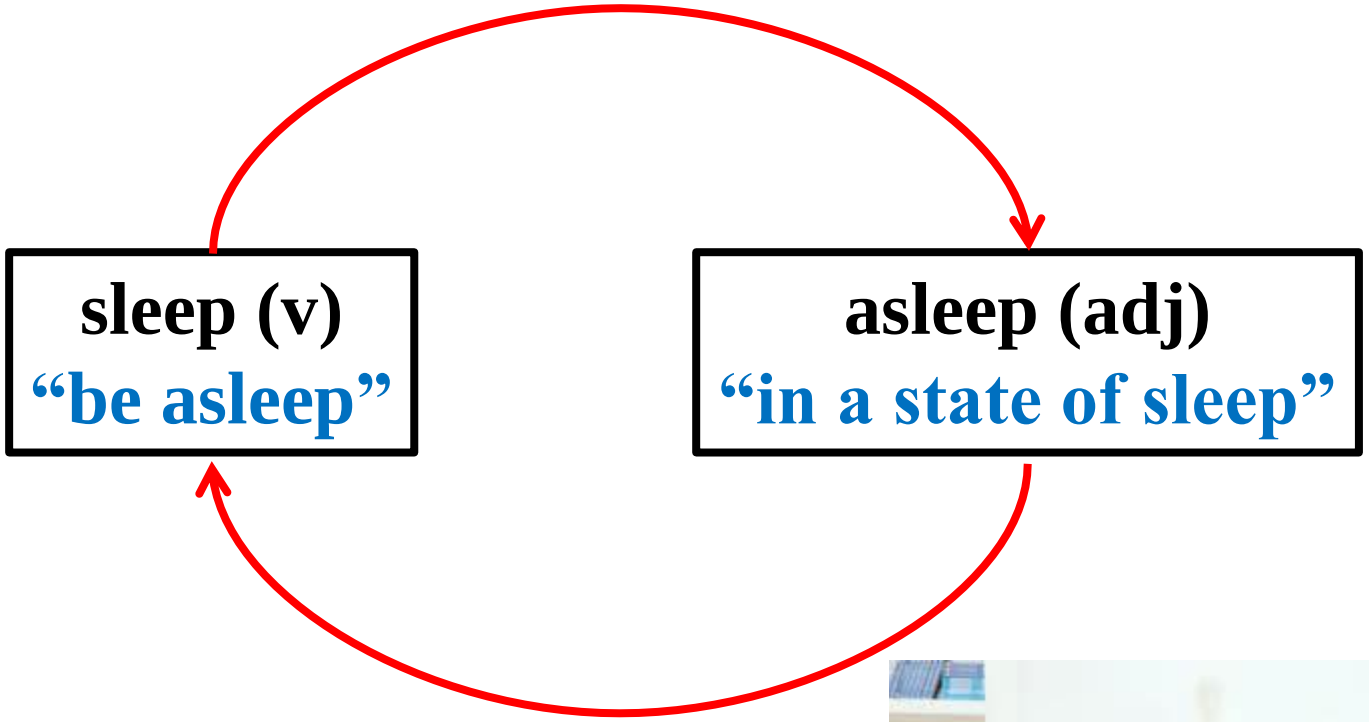
Department of Computer Science

University of Texas at Austin

Grounding Language Semantics in Perception and Action

- Most work in natural language processing deals only with text.
- The meaning of words and sentences is usually represented only in terms of other words or textual symbols.
- Truly understanding the meaning of language requires grounding semantics in perception and action in the world.

Sample Circular Definitions from WordNet



Historical Roots of Ideas on Language Grounding

- Meaning as Use & Language Games: Wittgenstein (1953)

- Symbol Grounding: Harnad (1990)

Direct Applications of Grounded Language

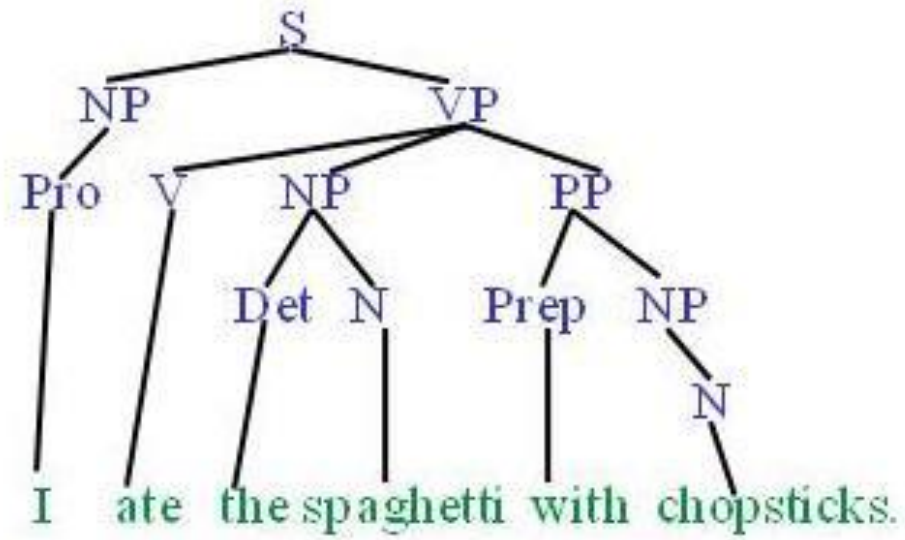
- Linguistic description of images and video
  - Content-based retrieval
  - Automated captioning for the visually impaired
  - Automated surveillance
- Human Robot Interaction
  - Obeying natural-language commands
  - Interactive dialog

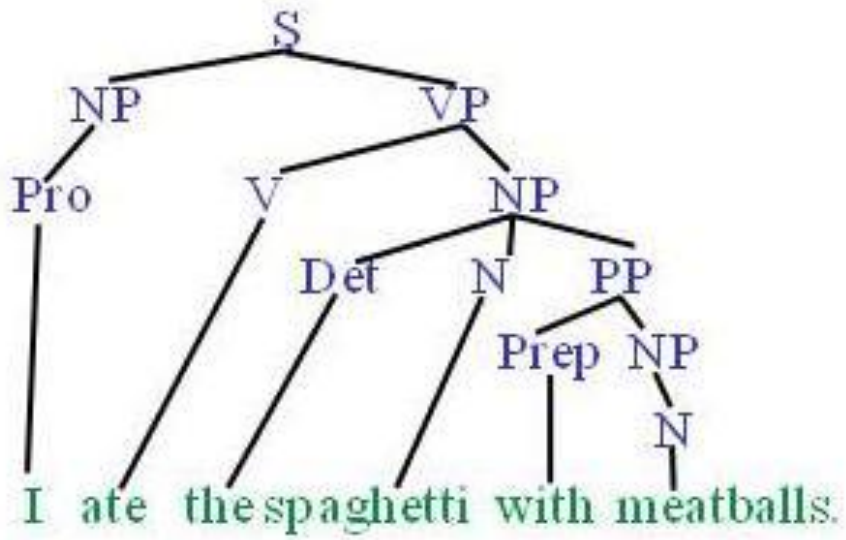
Supervised Learning and Natural Language Processing (NLP)

- Manual software development of robust NLP systems was found to be very difficult and inefficient.
- Most current state-of-the-art NLP systems are constructed by using machine learning methods trained on large supervised corpora.
  - POS-tagged text
  - Treebanks
  - Propbanks
  - Sense-tagged text

Syntactic Parsing of Natural Language

- Produce the correct syntactic parse tree for a sentence.





- Train and test on Penn Treebank with tens of thousands of manually parsed sentences.

Word Sense Disambiguation (WSD)

- Determine the proper dictionary sense of a word from its sentential context.
  - Ellen has a strong interest$_{\text{sense1}}$ in computational linguistics.
  - Ellen pays a large amount of interest$_{\text{sense4}}$ on her credit card.

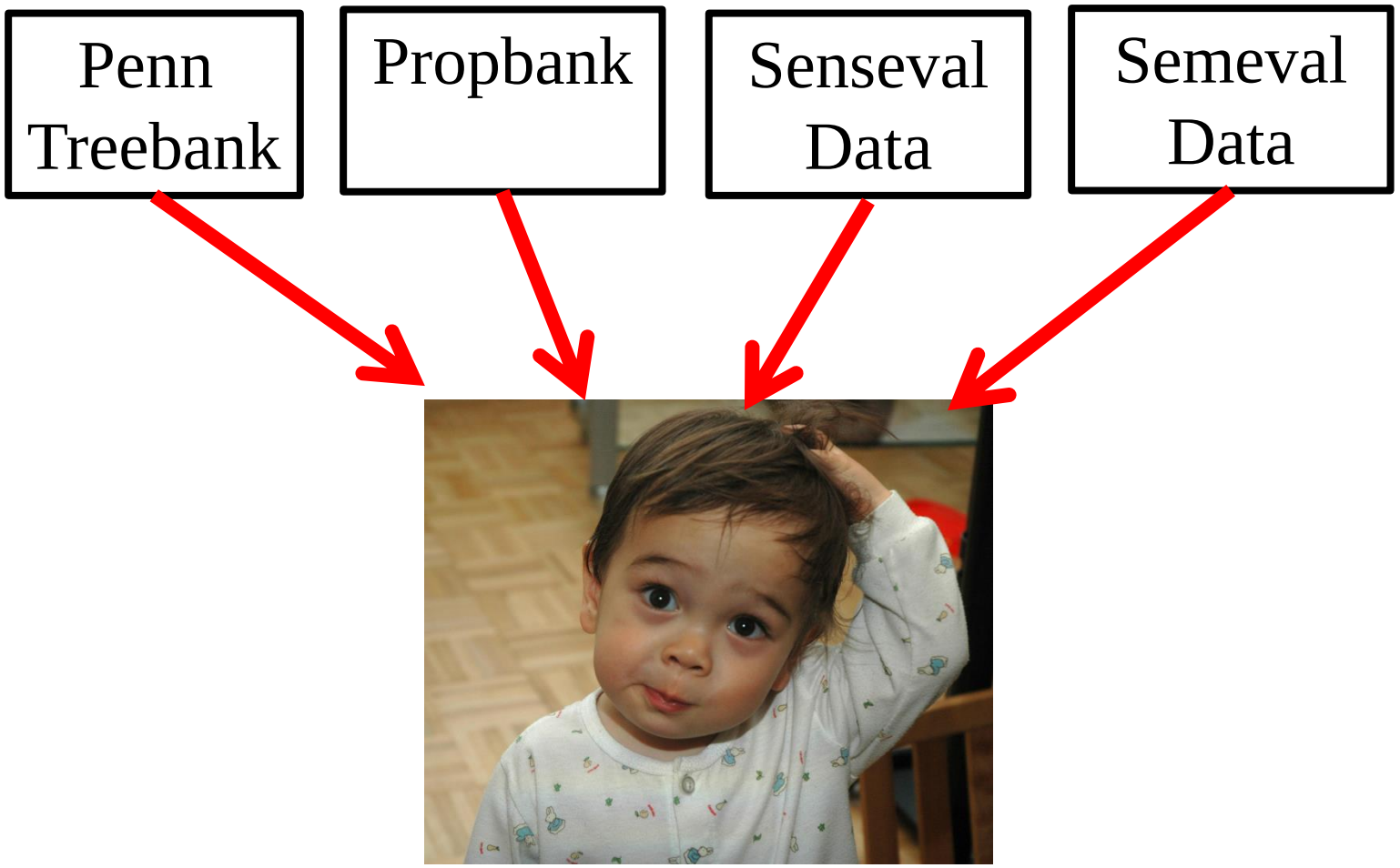
- Train and test on Senseval corpora containing hundreds of disambiguated instances of each target word.

Limitations of Supervised Learning

- Constructing supervised training data can be difficult, expensive, and time consuming.

- For many problems, machine learning has simply replaced the burden of knowledge and software engineering with the burden of supervised data collection.

Children do not Learn Language from Supervised Data

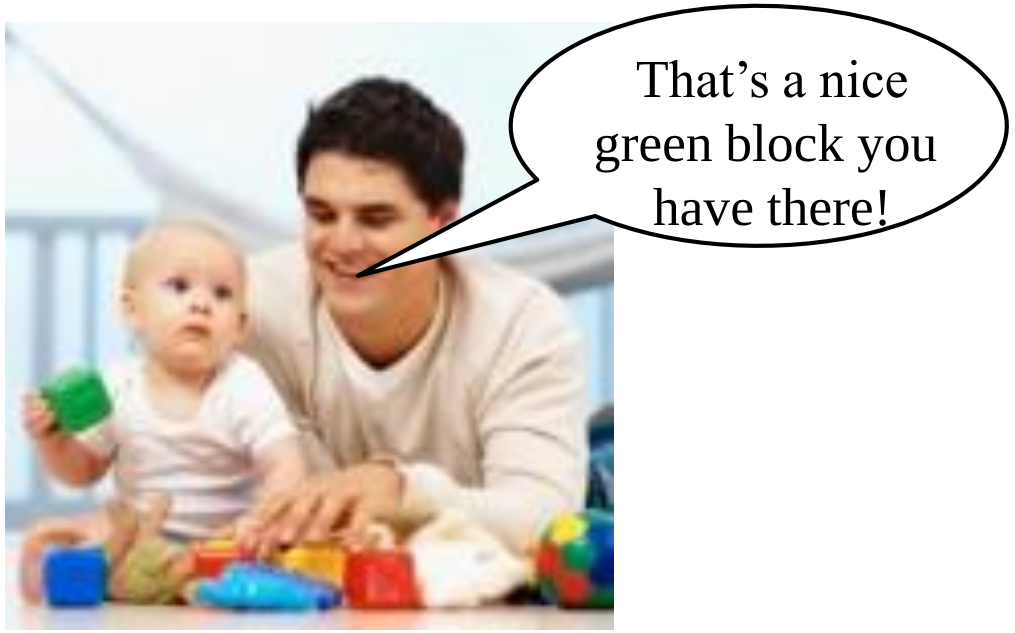


Children do not Learn Language from Raw Text

Unsupervised language learning is difficult and not an adequate solution since much of the requisite semantic information is not in the linguistic signal.

Learning Language from Perceptual Context

- The natural way to learn language is to perceive language in the context of its use in the physical and social world.
- This requires inferring the meaning of utterances from their perceptual context.



Grounded Language Learning in Virtual Environments

- Grounding in the real world requires sufficiently capable computer vision and robotics.
- Grounding in virtual environments is easier since perception and action are simulated.
- Given the prevalence of virtual environments (e.g. in games & education), linguistic communication with virtual agents also has practical applications.

Learning to Sportscast
(Chen, Kim, & Mooney, JAIR 2010)

- Learn to sportscast simulated Robocup soccer games by simply observing a person textually commentating them.

- Starts with ability to perceive events in the simulator, but no knowledge of the language.
- Learns to sportscast effectively in both English and Korean.

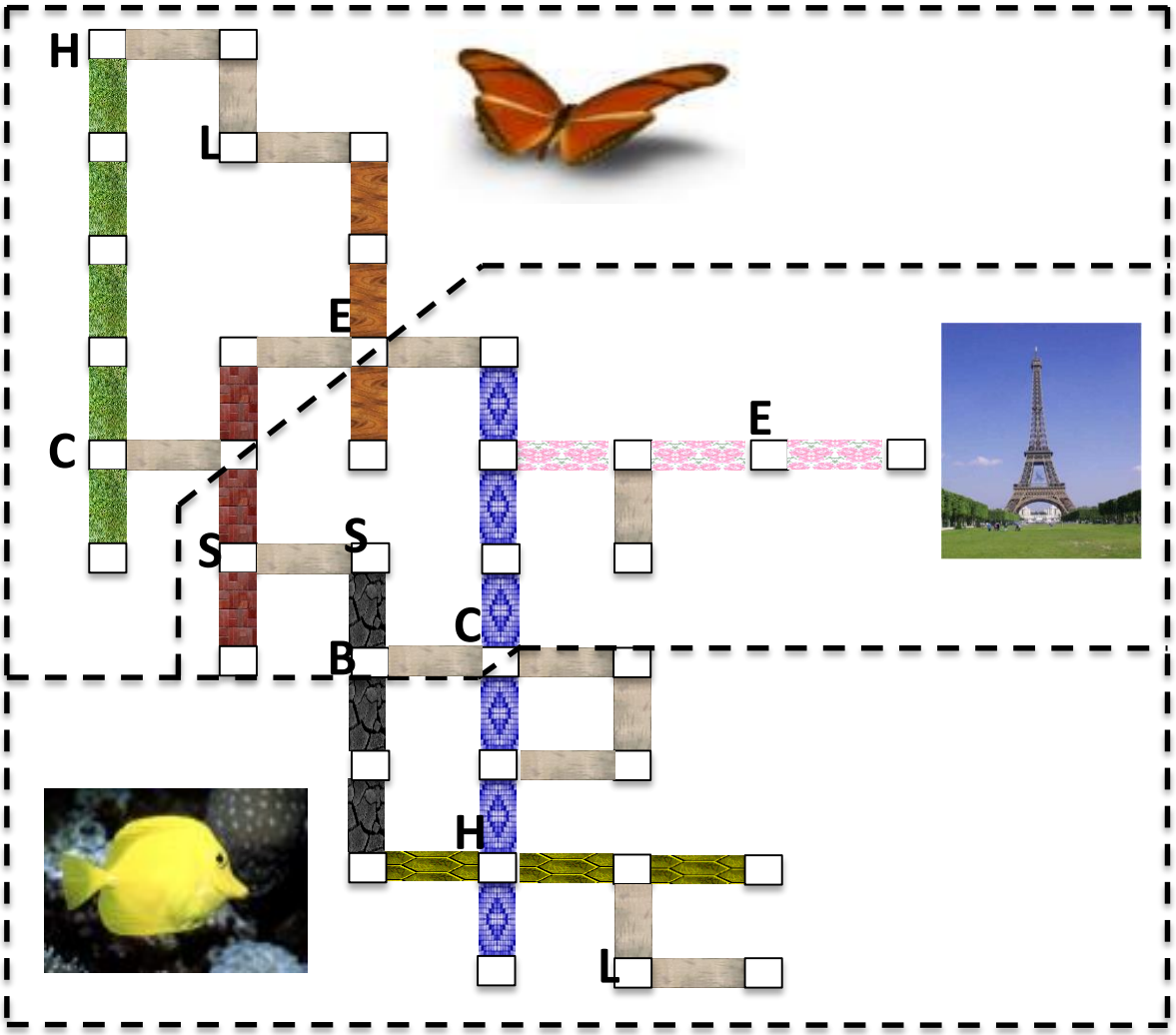
Machine Sportscast in English

Learning to Follow Directions in a Virtual Environment

- Learn to interpret navigation instructions in a virtual environment by simply observing humans giving and following such directions (Chen & Mooney, AAAI-11).

- **Eventual goal:** Virtual agents in video games and educational software that automatically learn to take and give instructions in natural language.

Sample Virtual Environment
(MacMahon, et al. AAAI-06)



H – Hat Rack

L – Lamp

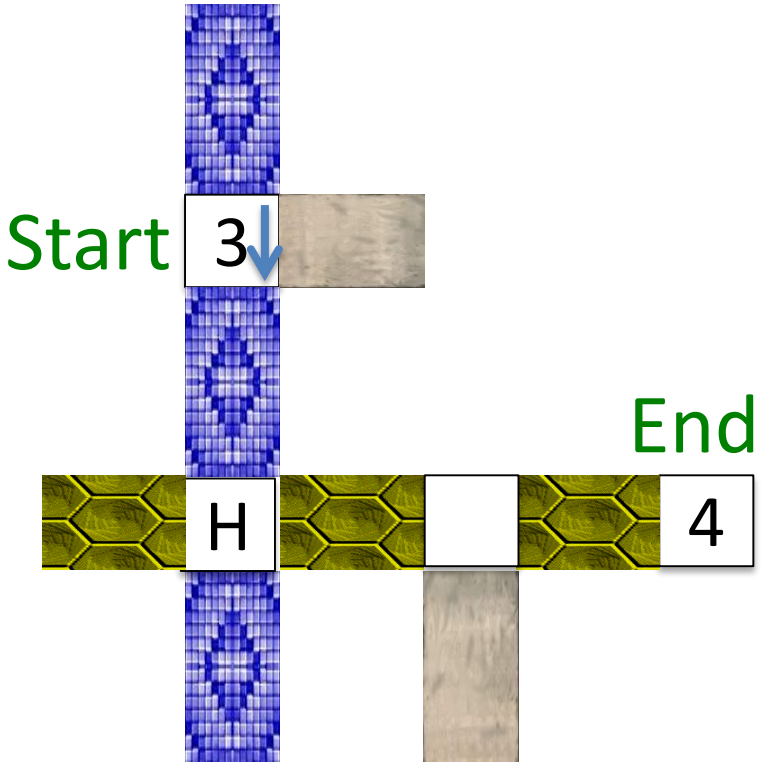
E – Easel

S – Sofa

B – Barstool

C - Chair

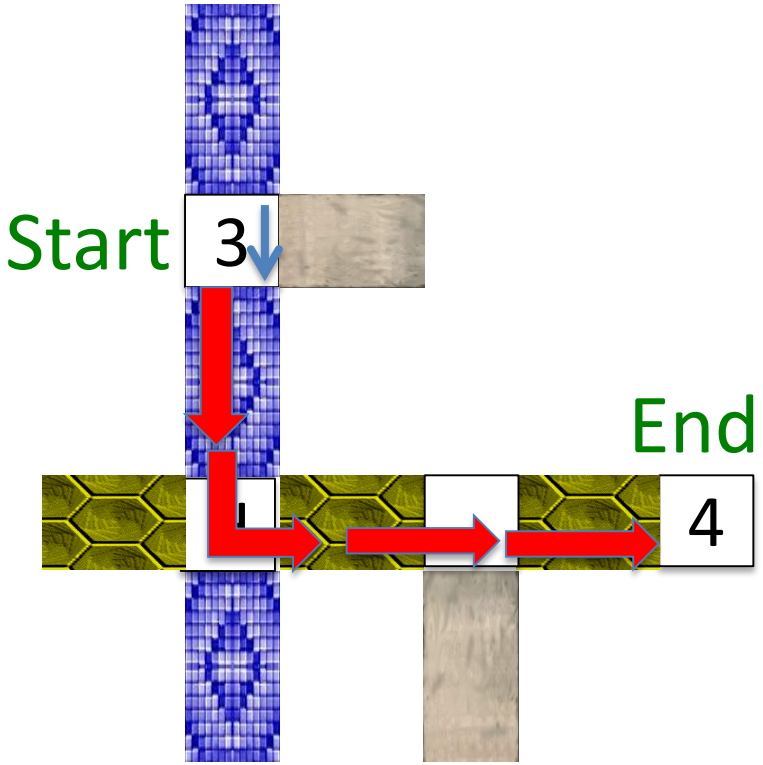
Sample Navigation Instructions



- Take your first left. Go all the way down until you hit a dead end.

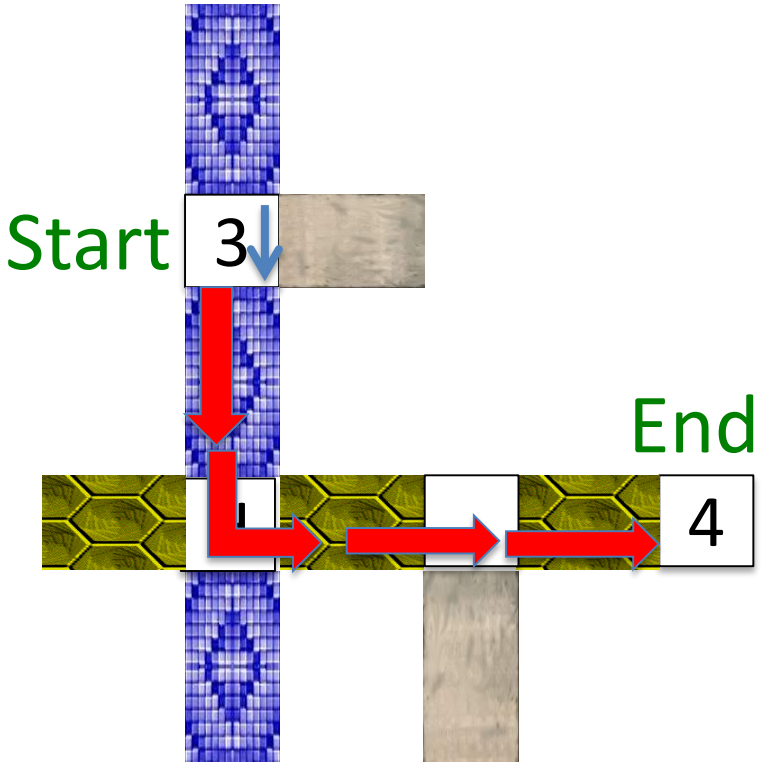
Sample Navigation Instructions

•Take your first left. Go all the way down until you hit a dead end.



Observed primitive actions:
Forward, Left, Forward, Forward

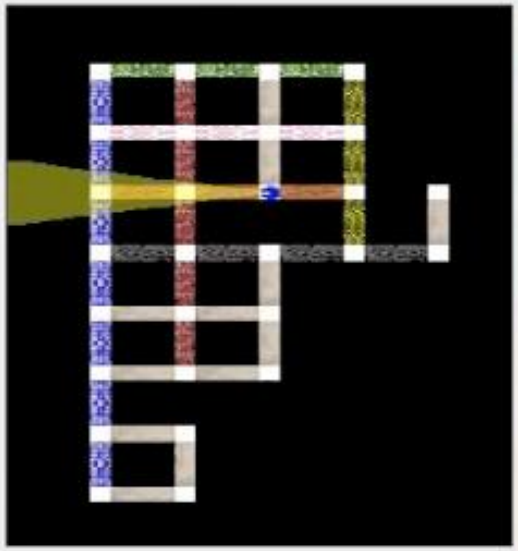
Sample Navigation Instructions



Observed primitive actions:
Forward, Left, Forward, Forward

- Take your first left. Go all the way down until you hit a dead end.
- Go towards the coat hanger and turn left at it. Go straight down the hallway and the dead end is position 4.
- Walk to the hat rack. Turn left. The carpet should have green octagons. Go to the end of this alley. This is p-4.
- Walk forward once. Turn left. Walk forward twice.

Observed Training Instance in Chinese



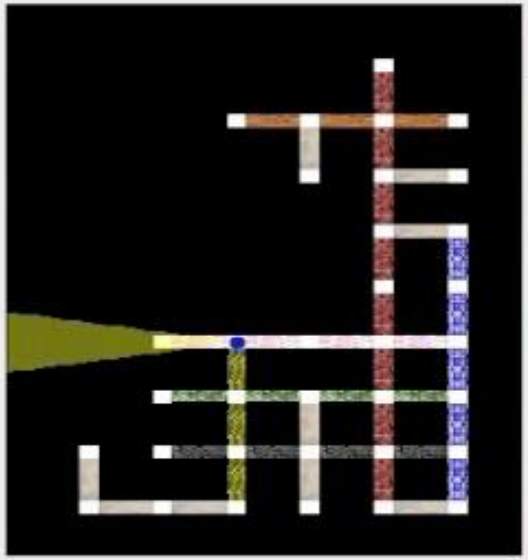
Simulation completion:

1X

Navigation Instruction

背靠 着 T 型 交界处 的 墙

Executing Test Instance in English
(after training in English)



Simulation completion:

1X

Navigation Instruction

Place your back against the wall of the 'T' intersection

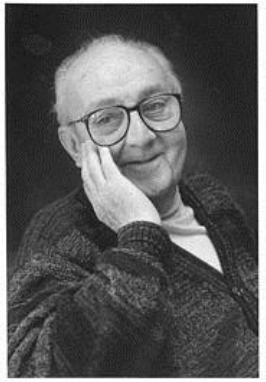
Parse

Turn(), Verify(back: WALL)

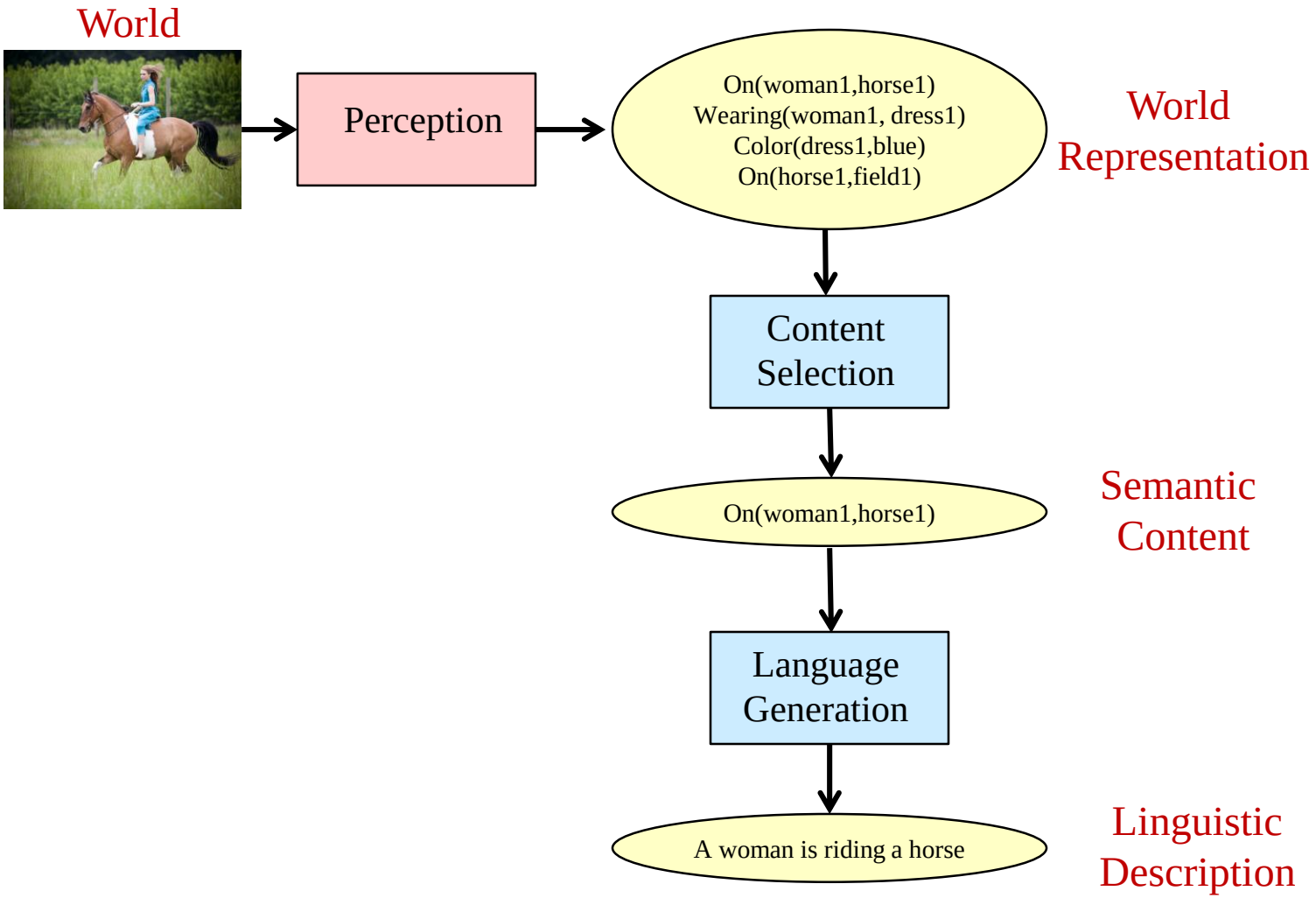
Statistical Learning and Inference for Grounded Language

- Use standard statistical methods to train a probabilistic model and make predictions.
- Construct a generative model that probabilistically generates language from observed situations.

George Box (1919-2013) :
"All models are wrong,
but some are useful."

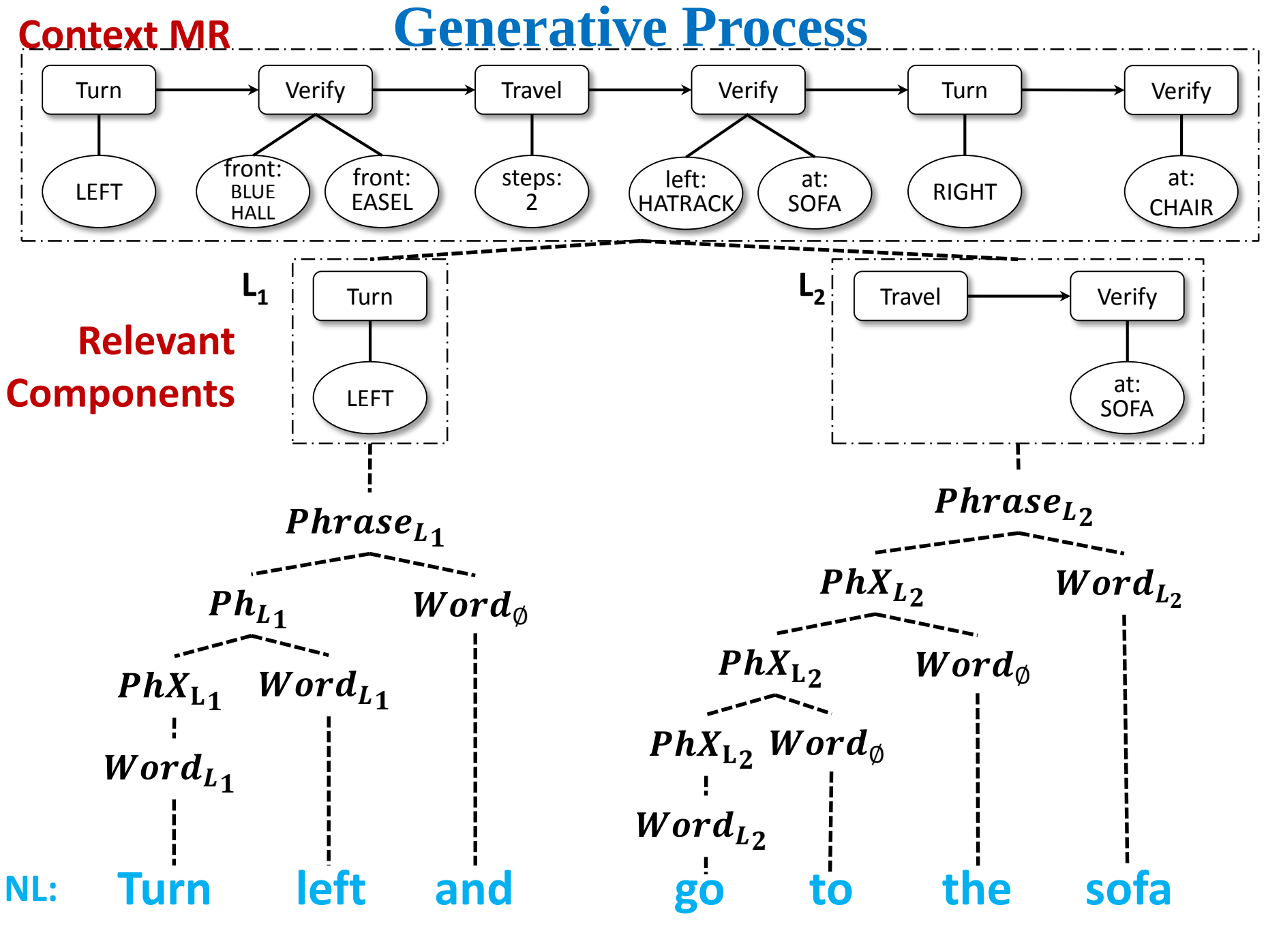


Probabilistic Generative Model for Grounded Language

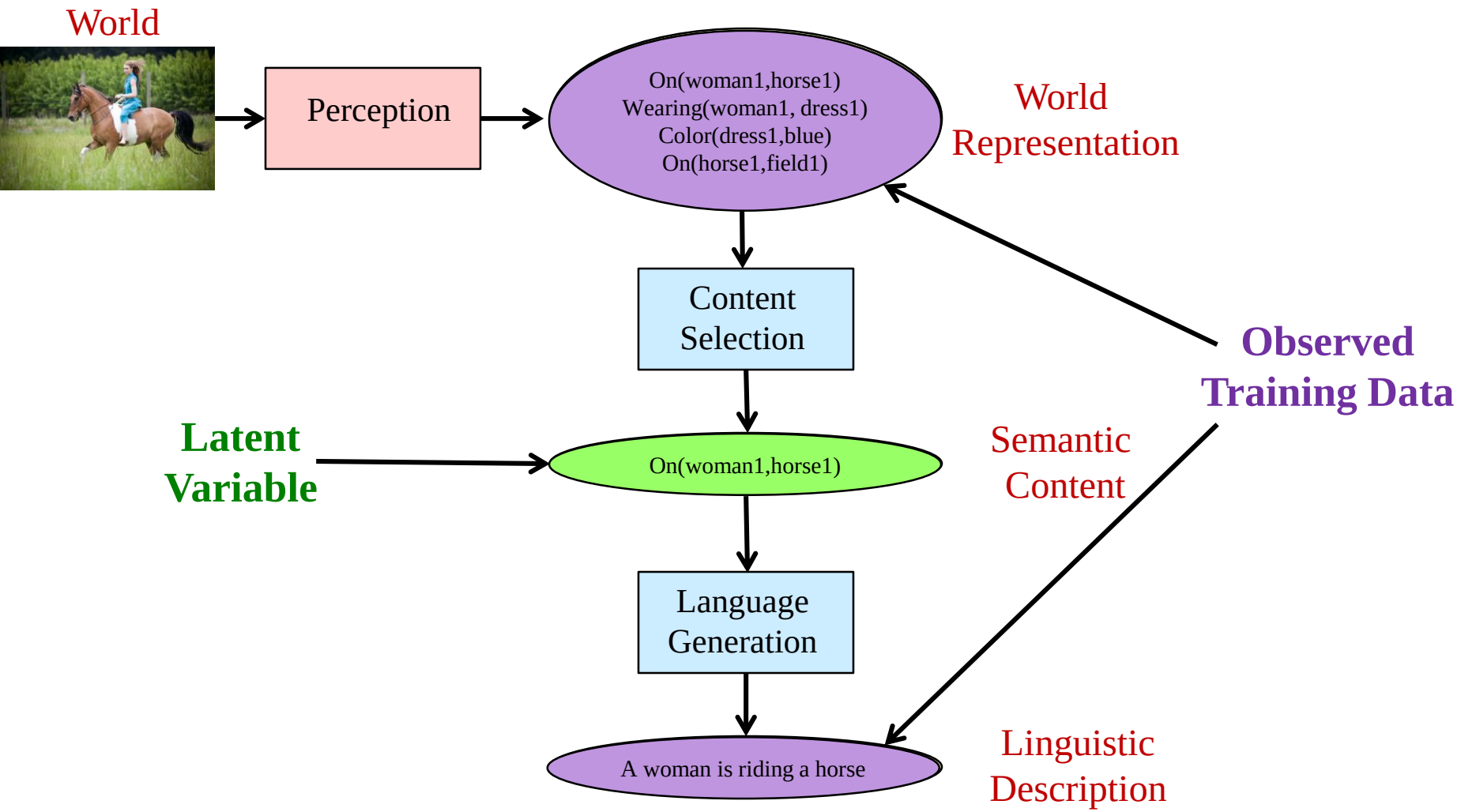


PCFGs for Grounded Language Generation

- Probabilistic Context-Free Grammars (PCFGs) can be used as a generative model for both content selection and language generation.
  - CFG with probabilistic choice of productions
- Initially demonstrated for Robocup sportscasting (Börschinger, Jones & Johnson, EMNLP-11).
- Later extended to navigation-instruction following by using prior semantic-lexicon learning (Kim & Mooney, EMNLP-12).


Context MR
Generative Process
Turn
Verify
Travel
Verify
Turn
Verify
LEFT
front: BLUE HALL
front: EASEL
steps: 2
left: HATRACK
at: SOFA
RIGHT
at: CHAIR
Relevant Components
L1
Turn
LEFT
L2
Travel
Verify
at: SOFA
Phrase_L1
Ph_L1
Word_∅
PhX_L1
Word_L1
Word_L1
Phrase_L2
PhX_L2
Word_L2
PhX_L2
Word_∅
PhX_L2
Word_∅
Word_L2
NL:
Turn
left
and
go
to
the
sofa

Generative Model Training for Grounded Language



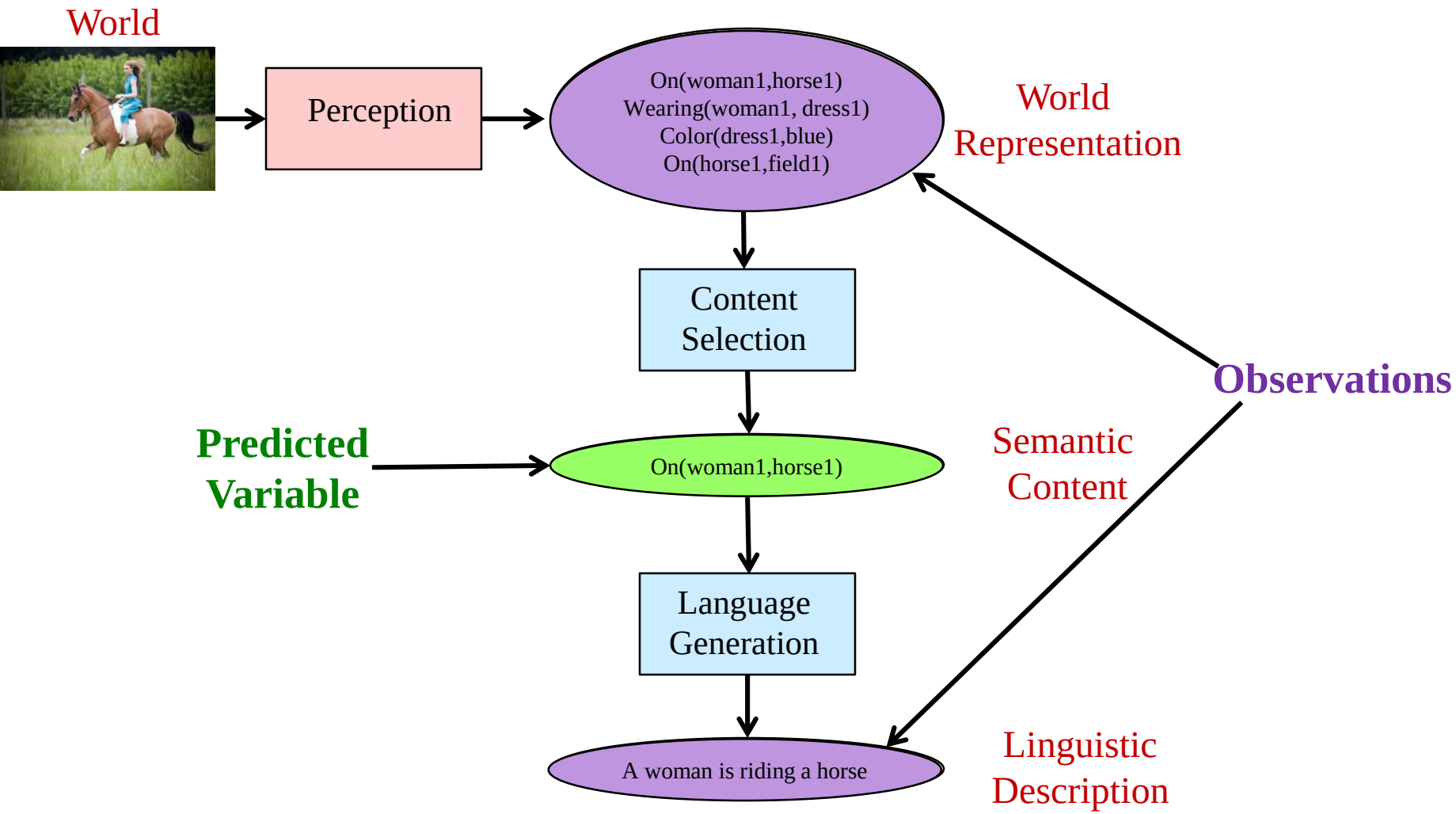
Statistical Training with Latent Variables

- Expectation Maximization (EM) is the standard method for training probabilistic models with latent variables.

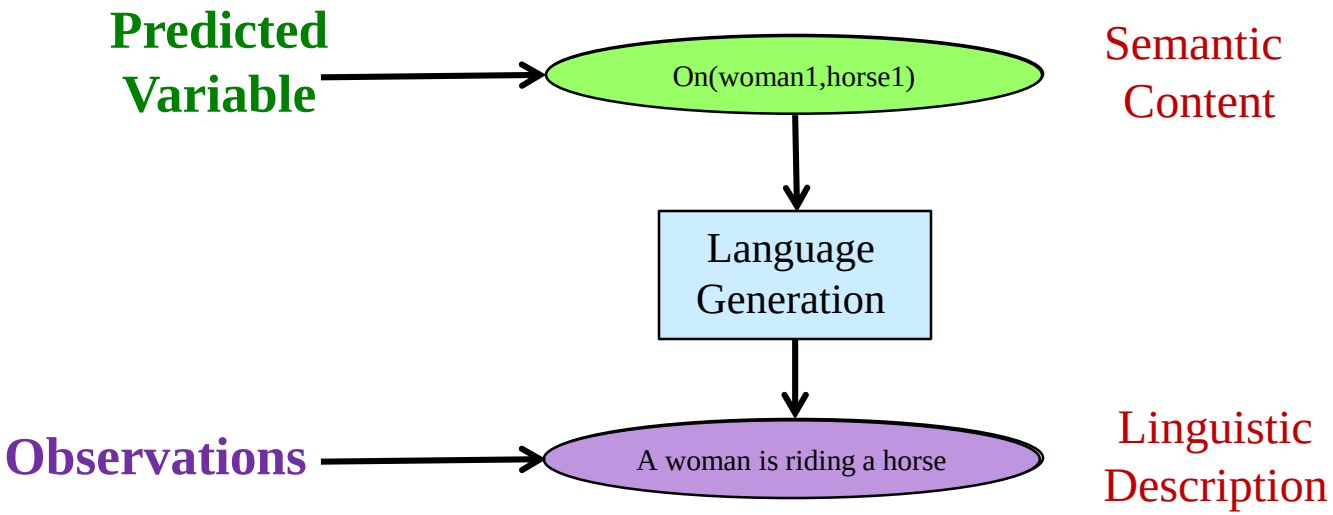
  Randomly initialize model parameters.
  Until convergence do:
  E Step: Compute the expected values of the latent variables given the observed data.
  M Step: Re-estimate the model parameters using these expected values and observed data.

- EM for PCFGs is call the Inside-Outside algorithm (Lari & Young, 1990).

Probabilistic Inference for Grounded Language



Probabilistic Inference for Grounded Language



Probabilistic Inference with Grounded PCFGs

- Determining the most probable parse of a sentence also determines its most likely latent semantic representation.
- An augmented version of the standard CYK CFG parsing algorithm can find the most probable parse in $O(n^3)$ time using dynamic programming.
  - Analogous to the Viterbi algorithm for a Hidden Markov Model (HMM)

Sample Successful Parse

Instruction: "Place your back against the wall of the 'T' intersection. Turn left. Go forward along the pink-flowered carpet hall two segments to the intersection with the brick hall. This intersection contains a hatrack. Turn left. Go forward three segments to an intersection with a bare concrete hall, passing a lamp. This is Position 5."

Parse: Turn ( ), Verify (back: WALL), Turn (LEFT), Travel ( ), Verify (side: BRICK HALLWAY), Turn (LEFT), Travel (steps: 3), Verify (side: CONCRETE HALLWAY)

Navigation-Instruction Following Evaluation Data

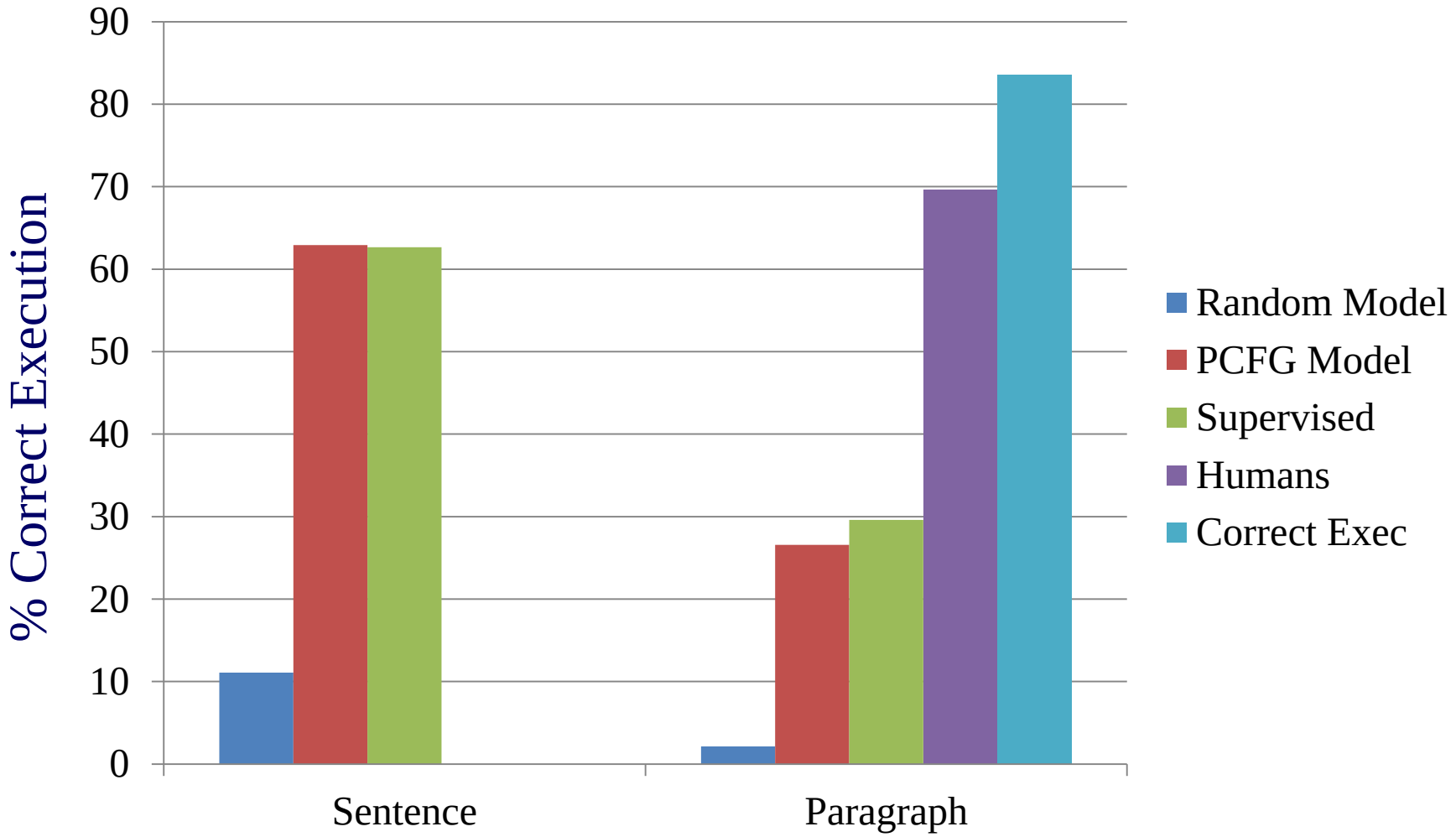
- 3 maps, 6 instructors, 1-15 followers/direction

	Paragraph	Single-Sentence
# Instructions	706	3,236
Avg. # sentences	5.0 (±2.8)	1.0 (±0)
Avg. # words	37.6 (±21.1)	7.8 (±5.1)
Avg. # actions	10.4 (±5.7)	2.1 (±2.4)

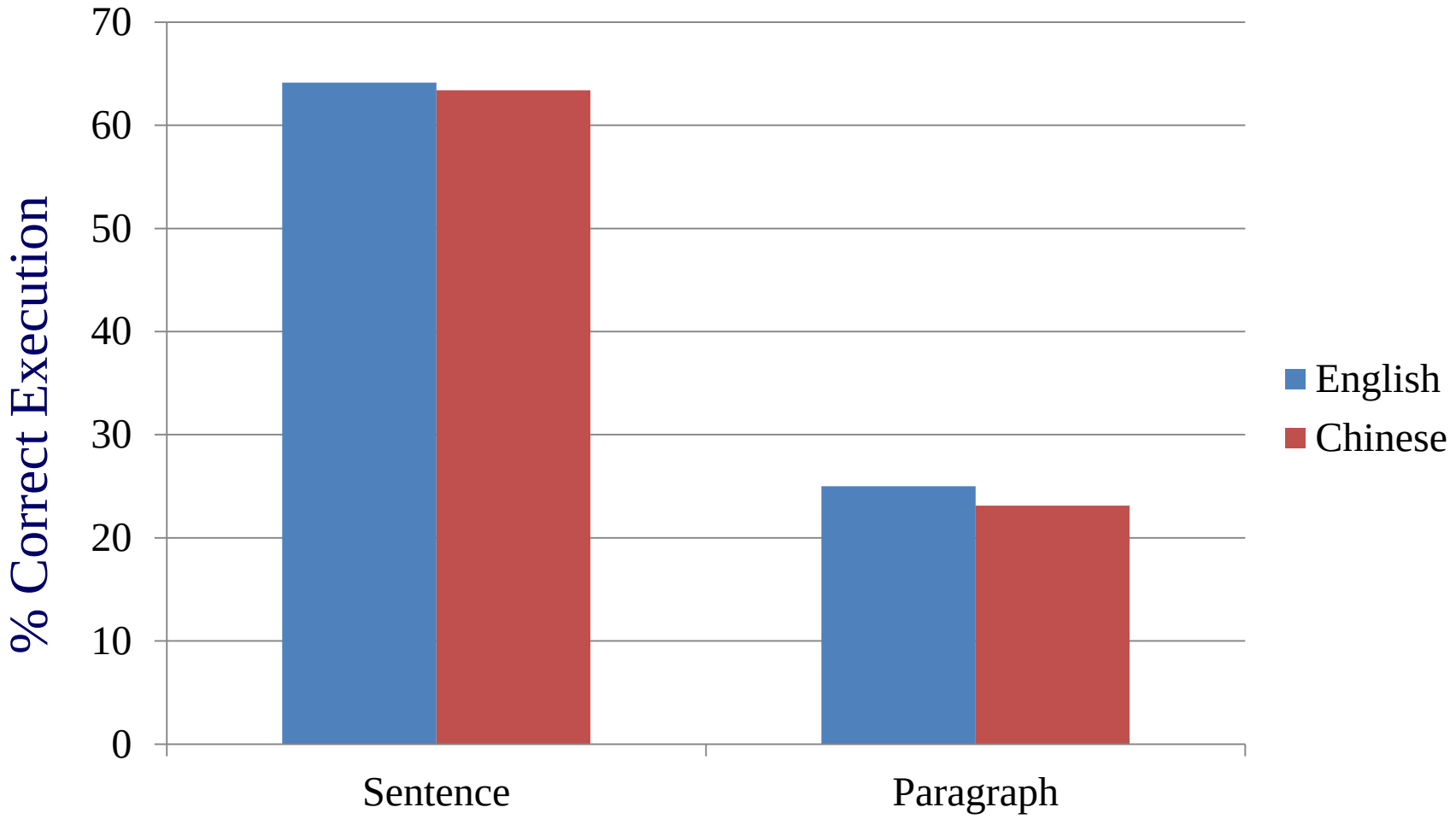
End-to-End Execution Evaluation

- Test how well the system follows new directions in novel environments.
  - Leave-one-map-out cross-validation.
- **Strict metric**: Correct iff the final position exactly matches goal location.
- **Lower baseline**:
  - Simple probabilistic generative model of executed plans without language.
- **Upper bounds**:
  - Supervised semantic parser trained on gold-standard plans.
  - Human followers.
  - Correct execution of instructions.

End-to-End Execution Results
English



End-to-End Execution Results
English vs. Mandarin Chinese



Grammar & Training Complexity

Data	|Grammar| # Productions	Time (hours)	EM Iterations
English	16,357	8.77	46
Chinese	15,459	8.05	40

Grounding in the Real World

- Move beyond grounding in simulated environments.

- Integrate NLP with computer vision and robotics to connect language to perception and action in the real world.

Grounded Language in Robotics

- Deb Roy at MIT has worked on grounded language for over a decade.
- He has developed a number of robots that learn and use grounded language.

Toco Robot from 2003

Real Robots You Can Instruct in Natural Language

- More recently, a group at MIT has developed a robotic forklift that obeys English commands (Tellex, et al., AAAI-11).
- Training data was collected in simulation using crowdsourcing on Amazon Mechanical Turk.
- Uses an existing English parser and direct semantic supervision to help learn to map sentences to formal robot commands.

Robotic Forklift NL Instruction Demo

Describing Pictures in Natural Language

- Several projects have explored automatically generating sentences that describe images.
- Typically trained on captioned/tagged images collected from the web, or crowdsourced human image descriptions.

(Rashtchian et al., 2010)

(1) A woman has a bird on her shoulder, and another bird on her head
(2) A woman with a bird on her head and a bird on her shoulder.
(3) A women sitting at a dining table with two small birds sitting on her.
(4) A young Asian woman sitting at a kitchen
table with a bird on her head and another on her shoulder.

Natural Language Generation for Images
(Kuznetsova et al., ACL-12)

- Trained on 1 million photos from Flickr that were filtered so that they contain useful captions.
- Extracts features from images using state-of-the-art object, scene, and “stuff” recognizers from computer vision.
- Composes sentences for novel images by using Integer Linear Programming to optimally stitch together phrases from similar training images.

Sample Generated Image Descriptions

ILP: This is a photo of this bird hopping around eating things off of the ground by river.

Human: IMG_6892 Lookn up in the sky its a bird its a plane its ah..... you

ILP: Taken in front of my cat sitting in a shoe box. Cat likes hanging around in my recliner.

Human: H happily rests his armpit on a warm Gatorade bottle of water (a small bottle wrapped in a rag)

This is ***a shoulder bag*** with a blended rainbow effect.

Generating English Descriptions for Videos

A person is riding a horse.

Video Description Research

- A few recent projects integrate visual object and activity recognition with NL generation to describe videos (Barbu et al., UAI-12, Khan & Gotoh, 2012).
- See our AAAI-13 talk:
  - Generating Natural-Language Video Descriptions Using Text-Mined Knowledge
    Niveda Krishnamoorthy

    Session 32A: NLP Generation and Translation

    11:50am, Thursday, July 18th

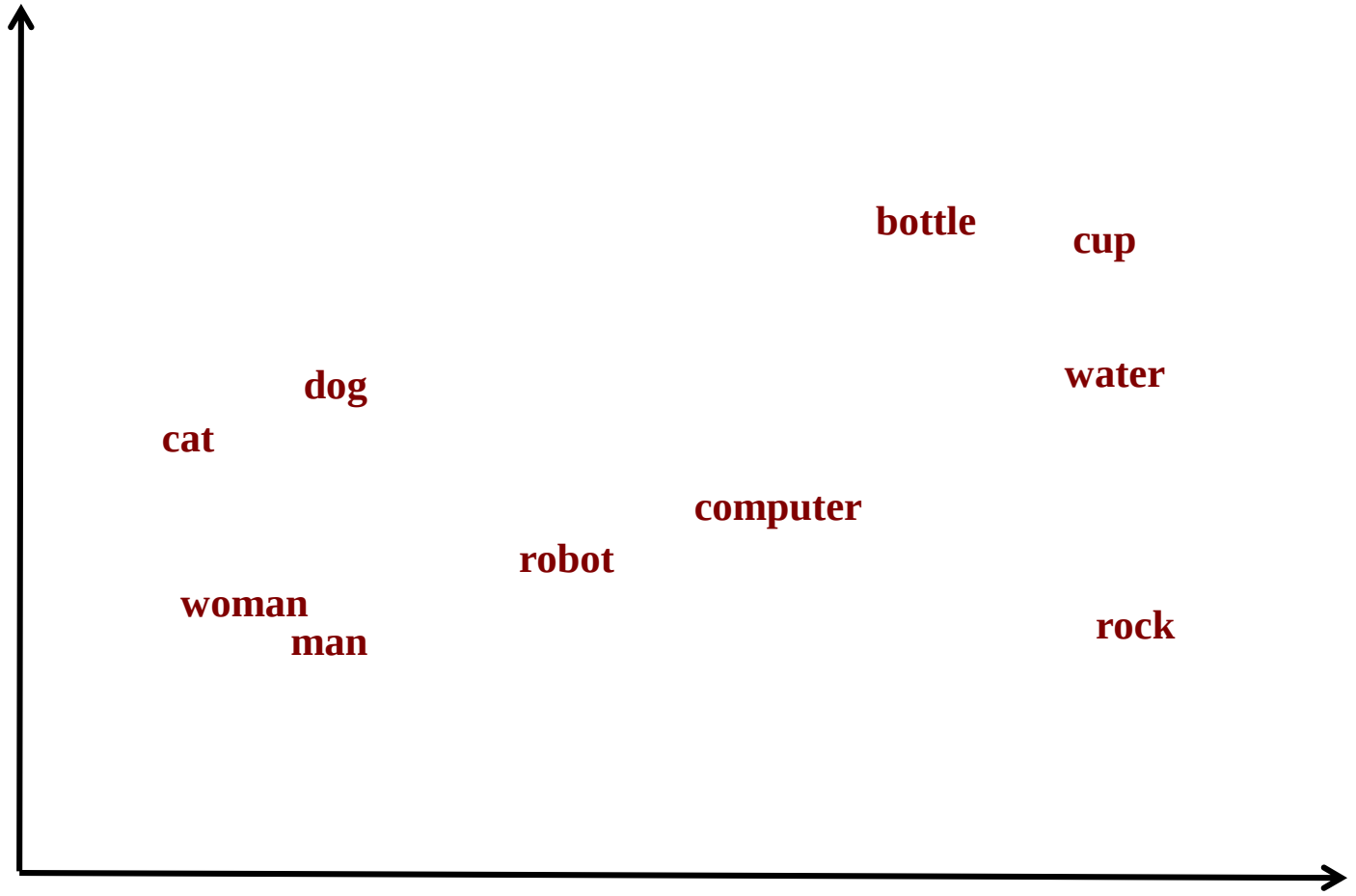
Connecting Word Meaning to Perception

- Word meanings as symbolic perceptual output
- Multimodal distributional semantics

Vector-Space (Distributional) Lexical Semantics

- Represent word meanings as points (vectors) in a (high-dimensional) Euclidian space.
- Dimensions encode aspects of the context in which the word appears (e.g. how often it co-occurs with another specific word).
  - "You will know a word by the company it keeps" (Firth)
- Semantic similarity defined as distance between points in this space.
- Many specific mathematical models for computing dimensions, dimensionality reduction, and similarity.
  - Latent Semantic Analysis (LSA)

Sample Lexical Vector Space (Reduced to Two Dimensions)



Multimodal Distributional Semantics

- Recent methods combine both linguistic and ***visual*** contextual features (Feng & Lapata, NAACL-10; Bruni et al., 2011; Silberer & Lapata, EMNLP-12) .
- Use corpus of captioned images to compute co-occurrence statistics between words and visual features extracted from images (e.g. color, texture, shape, detected objects).
- Multimodal models predict human judgments of lexical similarity better.
  - "cherry" more similar to "strawberry" than "orange"

Recent Spate of Workshops on Grounded Language

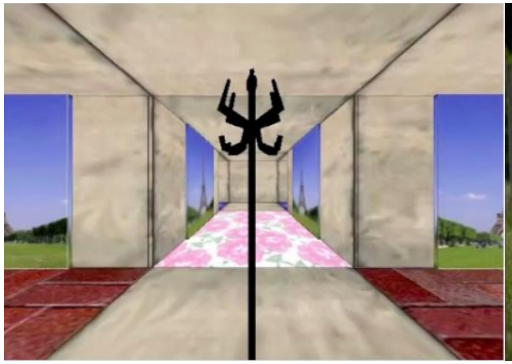
- AAAI-2011 Workshop on Language-Action Tools for Cognitive Artificial Agents: Integrating Vision, Action and Language
- NIPS-2011 Workshop on Integrating Language and Vision
- NAACL-2012 Workshop on Semantic Interpretation in an Actionable Context
- AAAI-2012 Workshop on Grounding Language for Physical Systems
- NAACL-2013 Workshop on Vision and Language
- CVPR-2013 Workshop on Language for Vision
- UW-MSR 2013 Summer Institute on Understanding Situated Language

Future Research Challenges

- Using linguistic and text-mined knowledge to aid computer vision.
- Active/interactive grounded language learning.
- Grounded-language dialog.
- Applications:
  - Language-enabled virtual agents
  - Language-enabled vision systems
  - Language-enabled robots

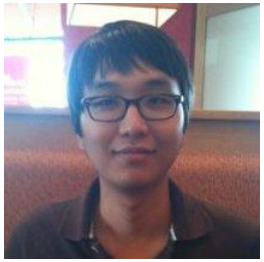
Conclusions

- Truly understanding language requires connecting it to perception and action.
- Learning from easily obtainable data in which language naturally co-occurs with perception and action improves NLP, vision, and robotics.
- The time is ripe to integrate language, vision, and robotics to address the larger AI problem.



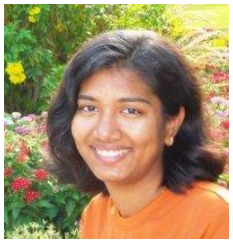
Thanks to
My (Former) Students and Colleagues!

David Chen



Joohyun Kim

Rohit Kate



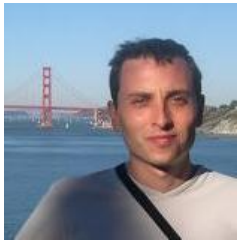
Sonal Gupta

Tanvi Motwani

Niveda Krishnamoorthy

Girish Malkarnenkar

Subhashini Venugopalan



Sergio Guadarrama

Kate Saenko

Kristen Grauman

Peter Stone