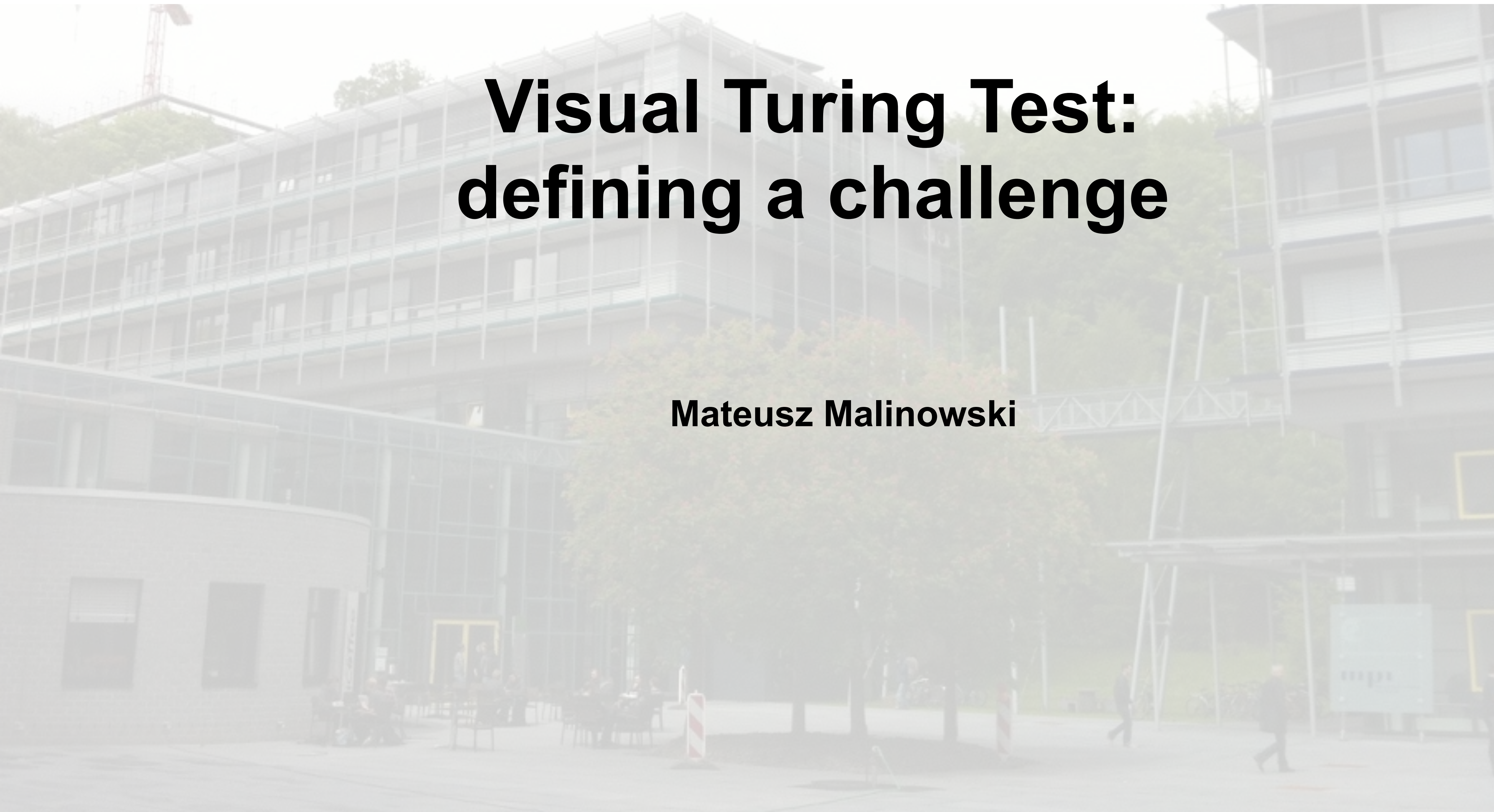
max planck institut
informatik



Visual Turing Test: defining a challenge

Mateusz Malinowski

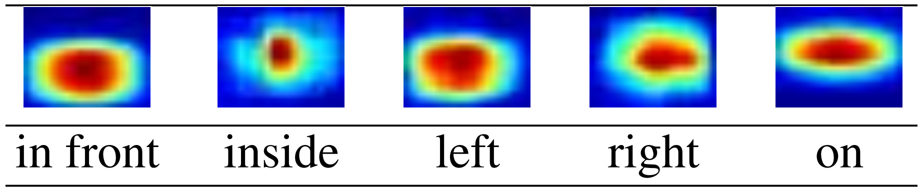
Visual Turing Test challenge

Ask about the content of the image

- How many sofas? → 3
- Where is the lamp? → on the table, close to tv
- What is behind the largest table? → tv
- What is the color of the walls? → purple

The task involves

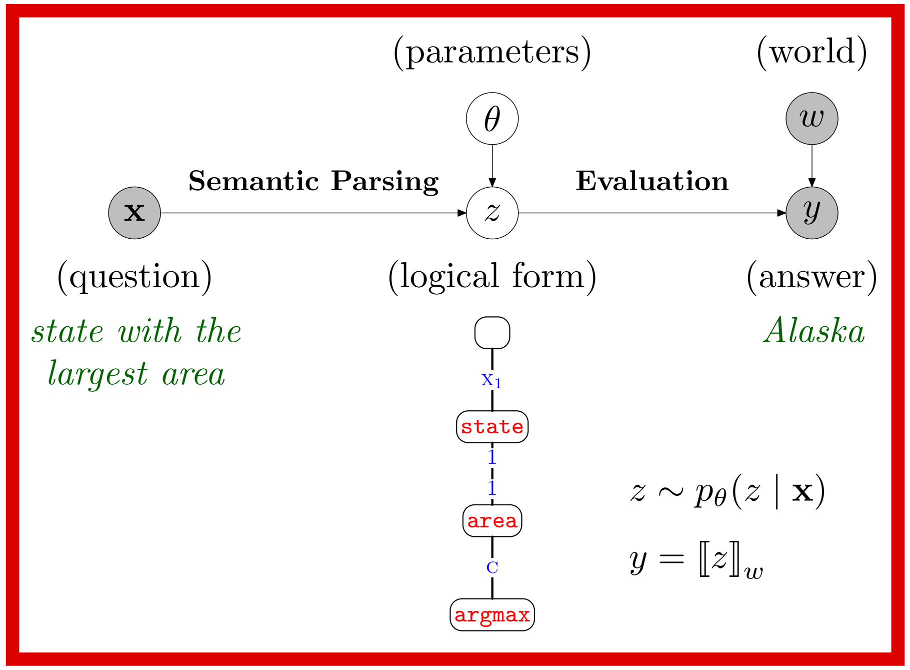
Object detection



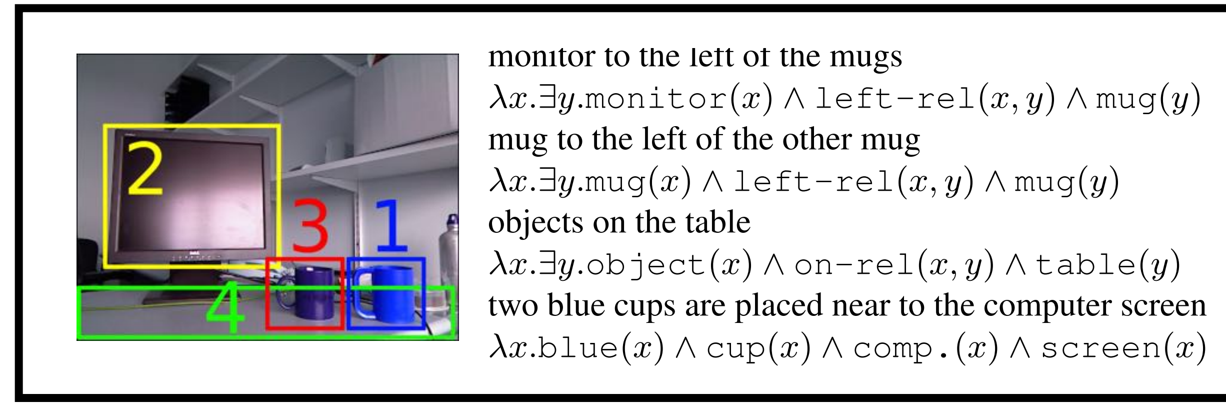
Spatial reasoning

Natural language understanding

Roadmap



Learning Dependency-Based Compositional Semantics
(P. Liang et. al. ACL 2011)

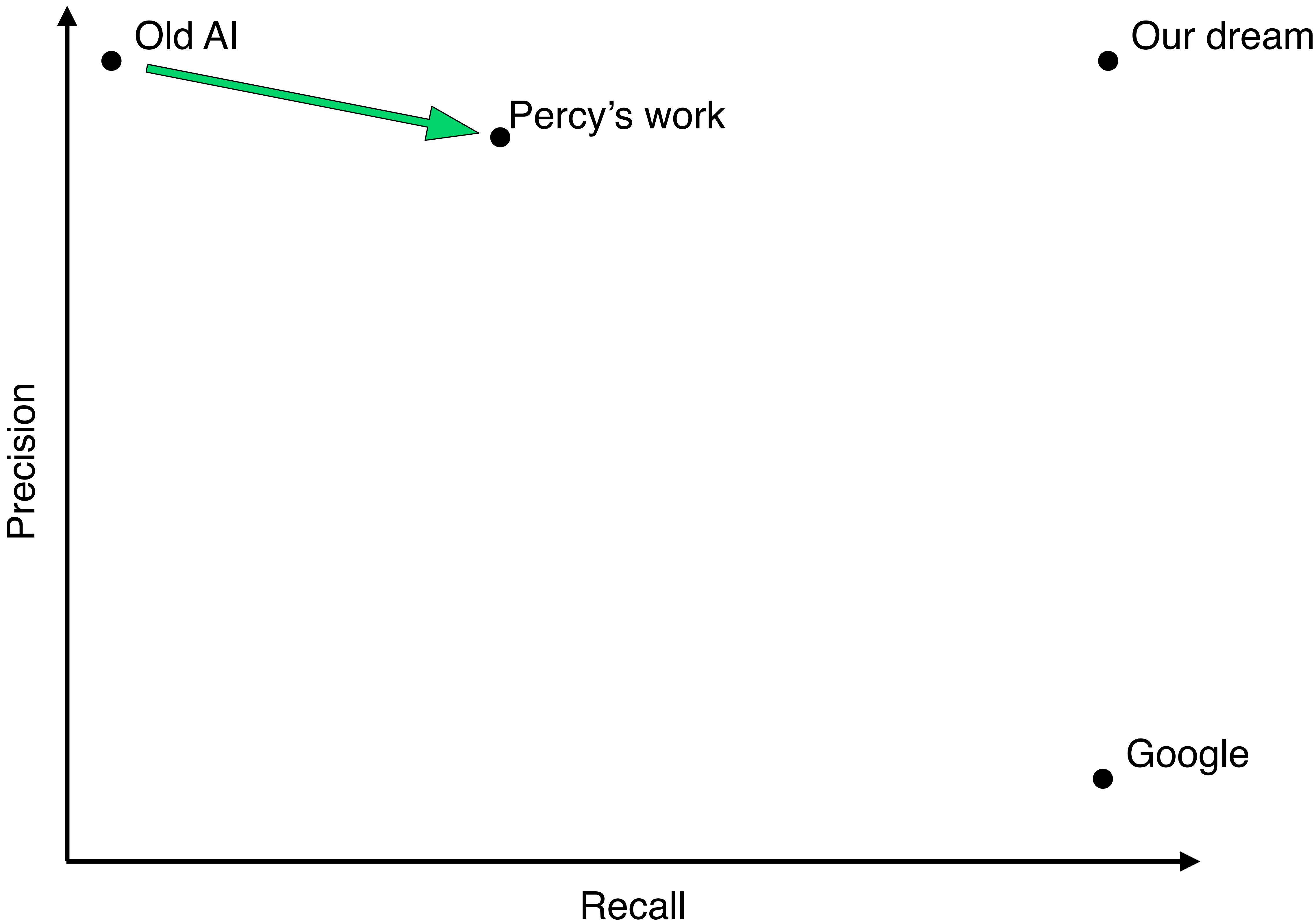


Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World.
(J. Krishnamurthy et. al. TACL 2013)

Some ideas

?

Two dimensions of language understanding



The Big Picture

What is the most populous city in California?

$\downarrow$

Database $\longrightarrow$ **System**

$\downarrow$

Los Angeles

Expensive: logical forms

[Zelle & Mooney, 1996; Zettlemoyer & Collins, 2005]

[Wong & Mooney, 2007; Kwiatkowski et al., 2010]

What is the most populous city in California?

$\Rightarrow \texttt{argmax}(\lambda x.\texttt{city}(x) \wedge \texttt{loc}(x, \texttt{CA}), \lambda x.\texttt{pop.}(x))$

How many states border Oregon?

$\Rightarrow \texttt{count}(\lambda x.\texttt{state}(x) \wedge \texttt{border}(x, \texttt{OR})$

...

Cheap: answers

[Clarke et al., 2010]

[**this work**]

What is the most populous city in California?

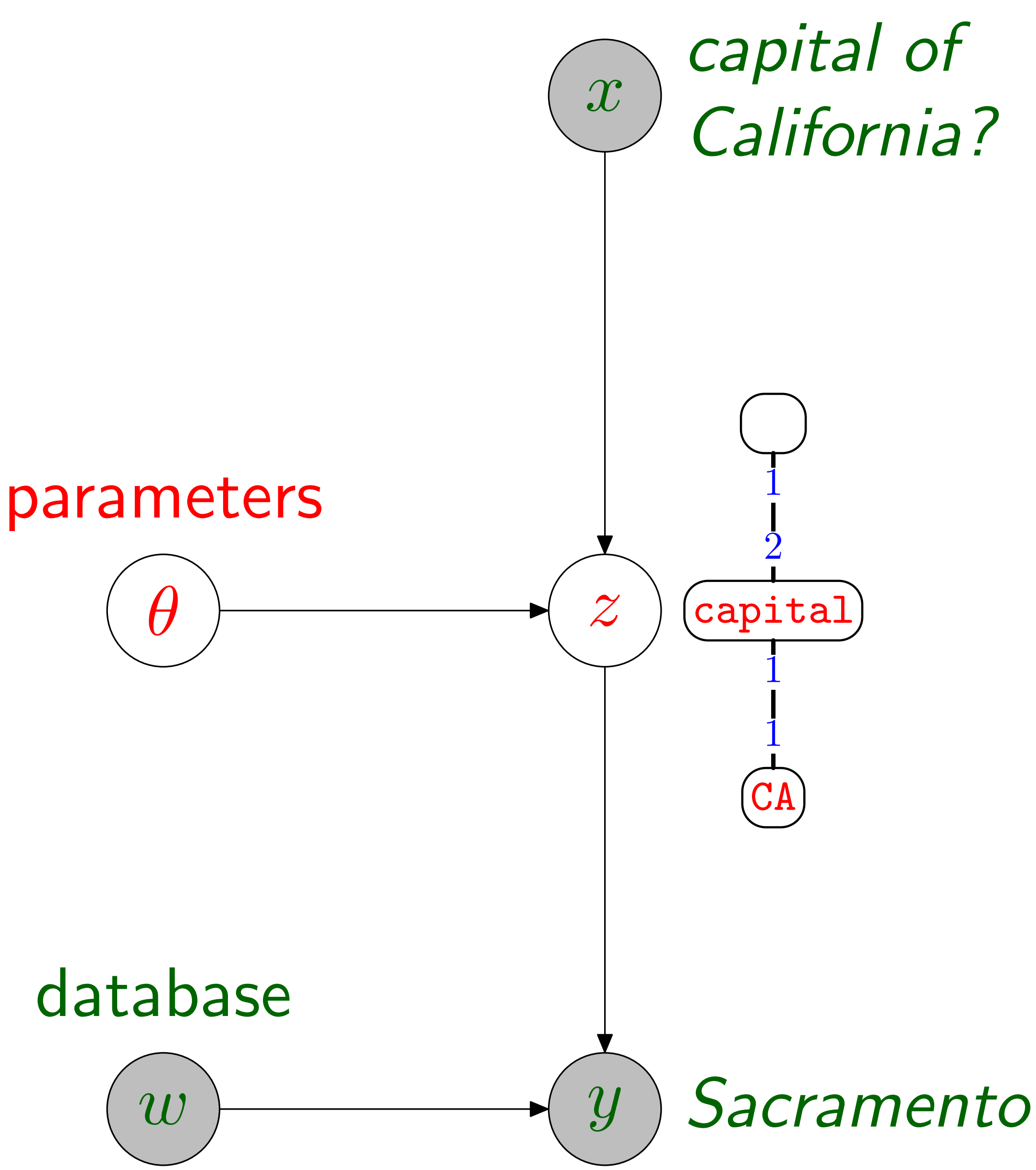
$\Rightarrow$ *Los Angeles*

How many states border Oregon?

$\Rightarrow$ *3*

...

The probabilistic framework



Interpretation $p(y \mid z, w)$

Semantic parsing $p(z \mid x, \theta)$

Objective

$$\max_{\theta} \sum_{z} \underbrace{p(y \mid z, w)}_{\textbf{Interpretation}} \underbrace{p(z \mid x, \theta)}_{\textbf{Semantic parsing}}$$

Learning

parameters θ

$(0.2, -1.3, \ldots, 0.7)$ $\xrightarrow{\text{enumerate/score DCS trees}}$ k-best list

$\xleftarrow[\text{numerical optimization (L-BFGS)}]{}$

k-best list	
tree1	✗
tree2	✗
tree3	✓
tree4	✗
tree5	✗

Challenges of the semantic parsing

What is the most populous city in California?

$$\lambda x.\texttt{city}(x) \wedge \texttt{loc}(x, \texttt{CA})$$

Los Angeles

↓

What is the most populous city in California?

$$\lambda x.\texttt{state}(x) \wedge \texttt{border}(x, \texttt{CA})$$

Los Angeles

↓

What is the most populous city in California?

$$\texttt{argmax}(\lambda x.\texttt{city}(x) \wedge \texttt{loc}(x, \texttt{CA}), \lambda x.\texttt{population}(x))$$

Los Angeles

Challenges of the semantic parsing

Words to Predicates (Lexical Semantics)

				`city`	`city`			
				`state`	`state`			
				`river`	`river`			
			`argmax`	`population`	`population`		`CA`	
What	*is*	*the*	*most*	*populous*	*city*	*in*	*CA*	*?*

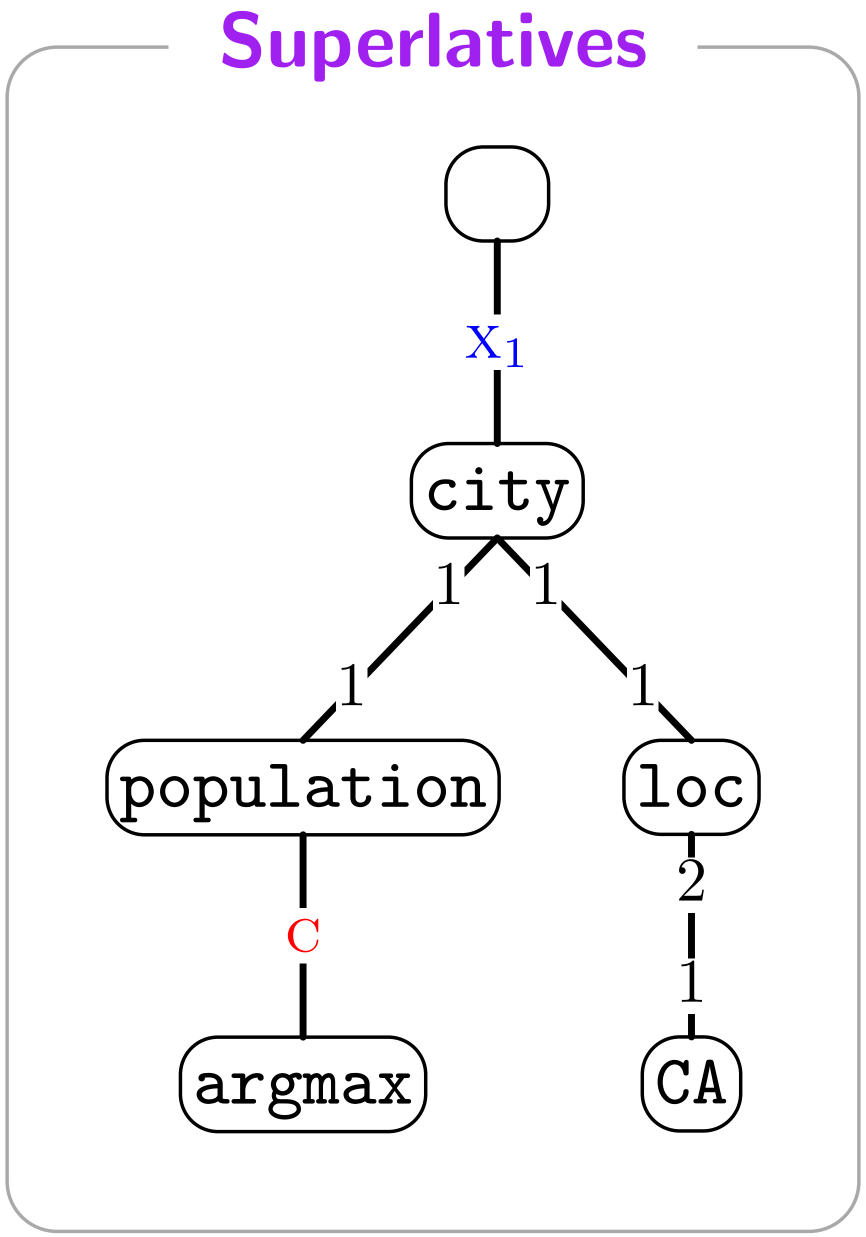
Lexical Triggers:

1. String match *CA* ⇒ `CA`
2. Function words (20 words) *most* ⇒ `argmax`
3. Nouns/adjectives *city* ⇒ `city state river population`

Dependency-based compositional semantics

Solution: Mark-Execute

most populous city in California

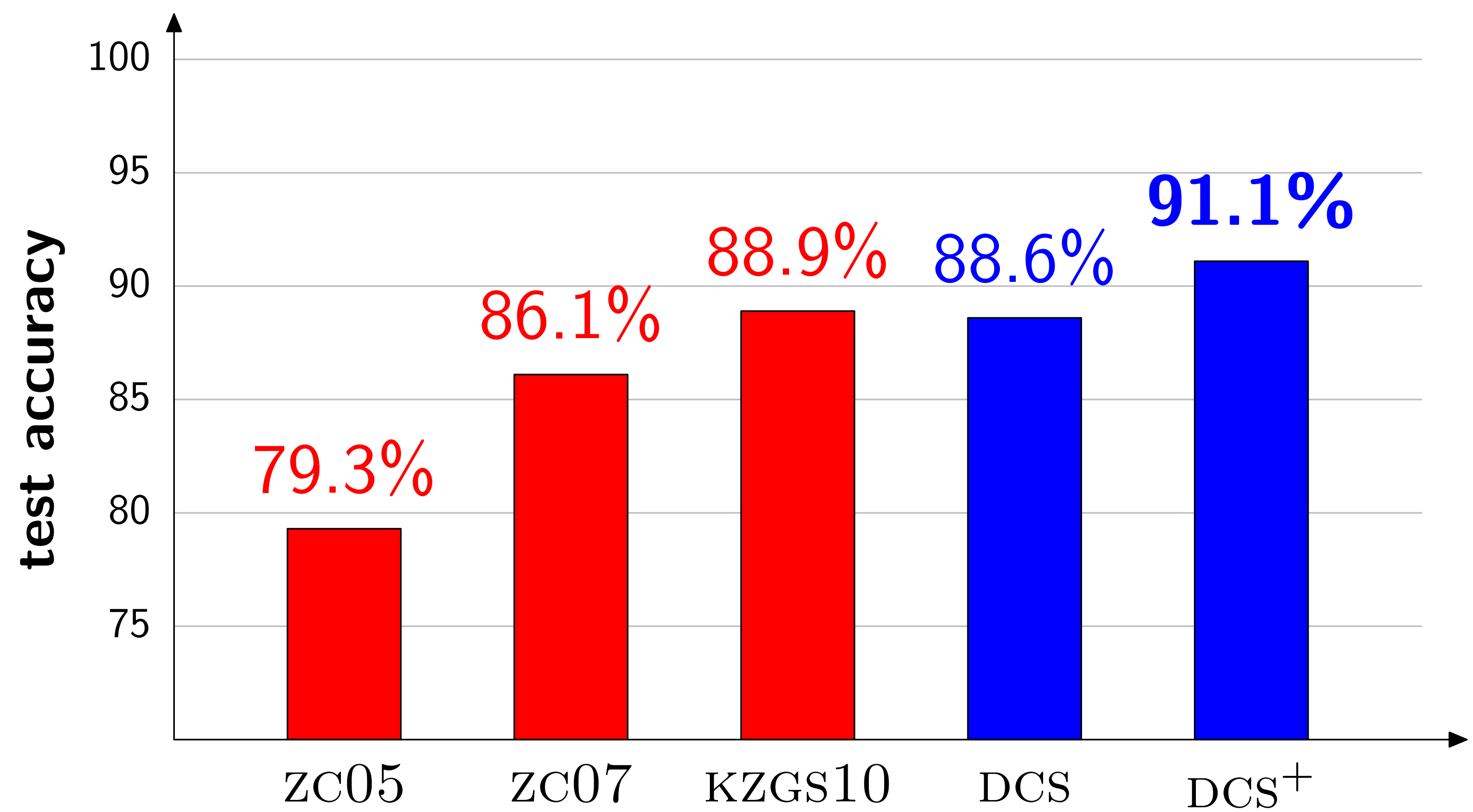


Mark at syntactic scope

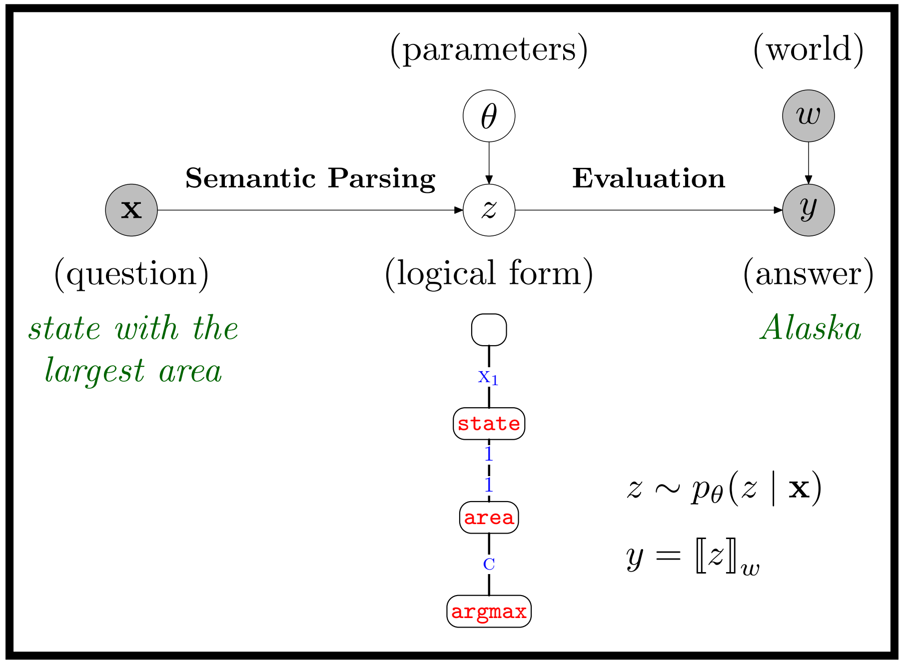
Results

On GEO, 600 training examples, 280 test examples

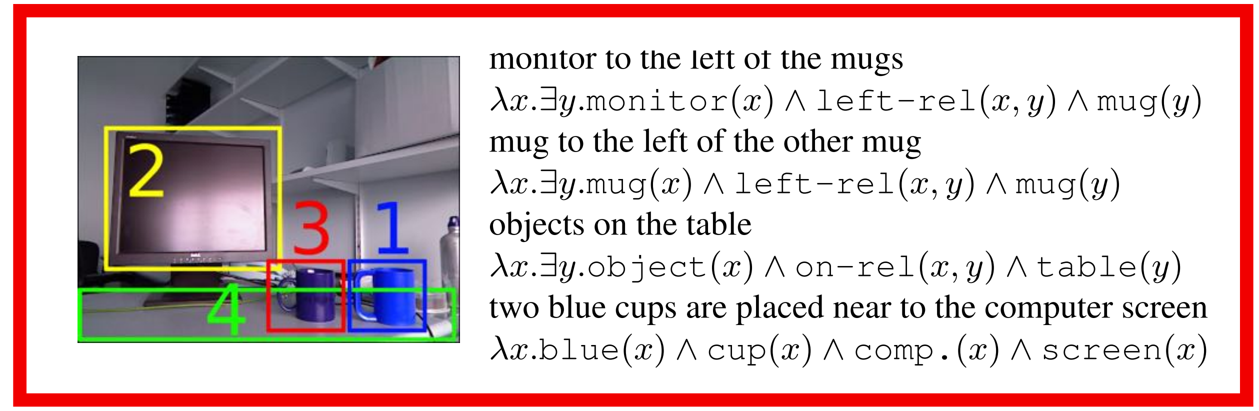
System	Description	Lexicon	Logical forms
ZC05	CCG [Zettlemoyer & Collins, 2005]	✗ ✗	✓
ZC07	relaxed CCG [Zettlemoyer & Collins, 2007]	✗ ✗	✓
KZGS10	CCG w/unification [Kwiatkowski et al., 2010]	✗ ✗	✓
DCS	our system	✓ ✗	✗
DCS$^+$	our system	✓ ✓	✗



Roadmap



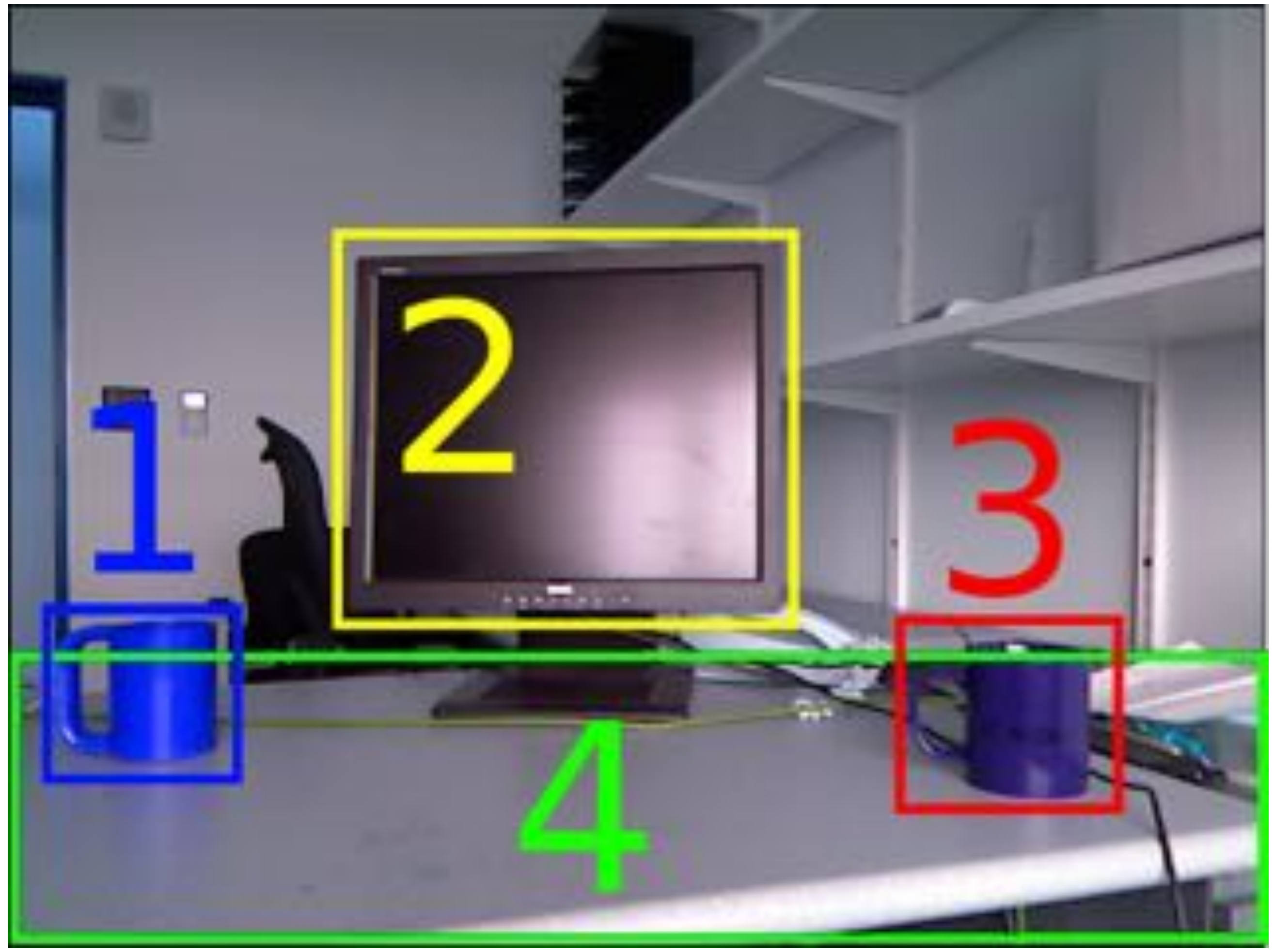
Learning Dependency-Based Compositional Semantics
(P. Liang et. al. ACL 2011)



Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World.
(J. Krishnamurthy et. al. TACL 2013)

Some ideas

Grounding problem

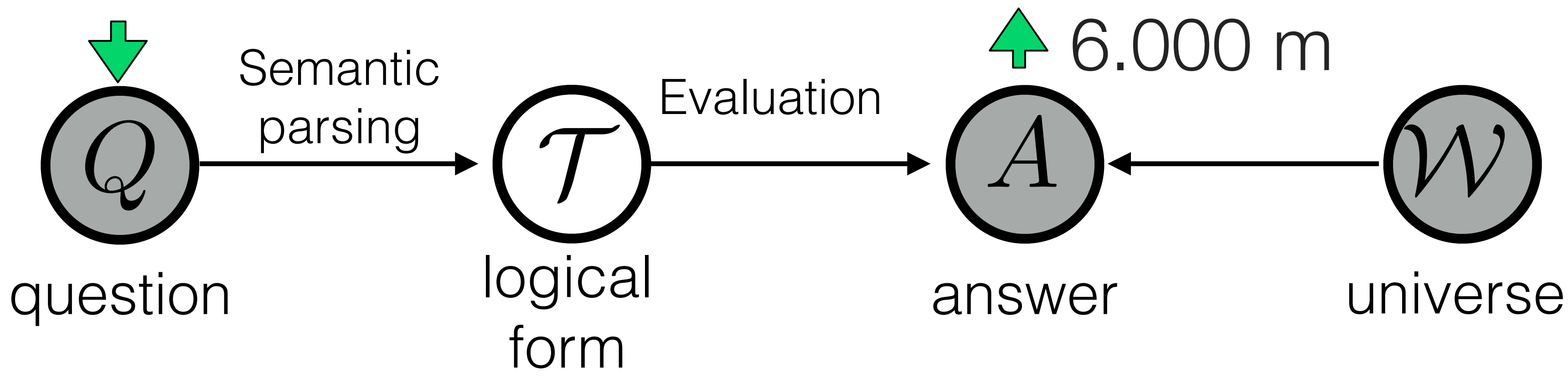


The mugs $=$ {  ,  }

A mug left of the monitor $=$ {  }

Question answering problem

How high is the highest point in the largest state?

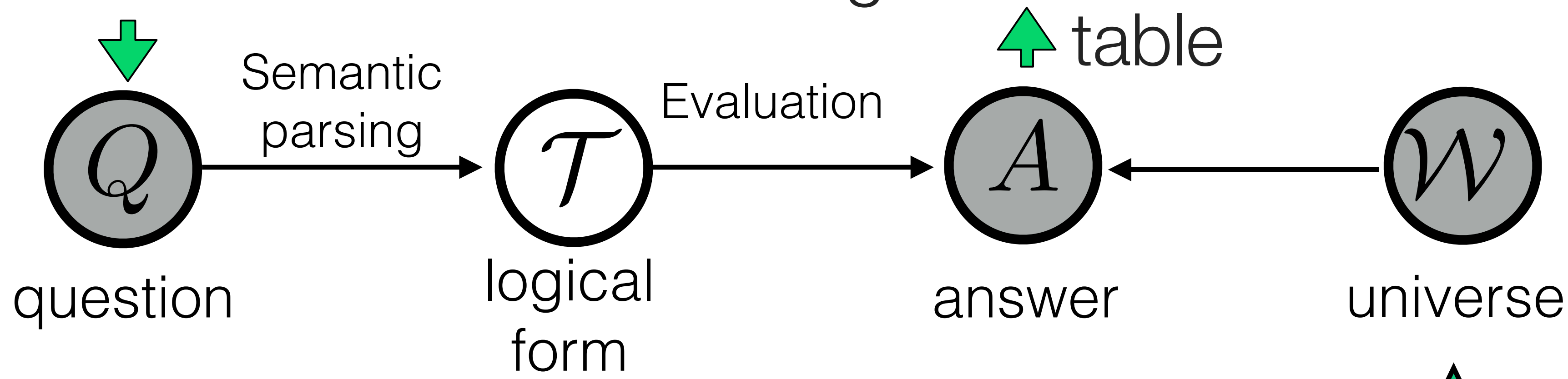


P. Liang, M. Jordan, D. Klein. Learning Dependency-Based Compositional Semantics. ACL'11

J. Berant, A. Chou, R. Frostig, and P. Liang. Semantic Parsing on Freebase from Question-Answer Pairs. EMNLP'13.

Question answering problem

What is in front of sofa in image 1?



Our knowledge base

Scene analysis

sofa (1,brown, image 1, X,Y,Z)

table(1,brown, image 1,X,Y,Z)

wall (1,white, image 1, X,Y,Z)

bed (1, white, image 2 X,Y,Z)

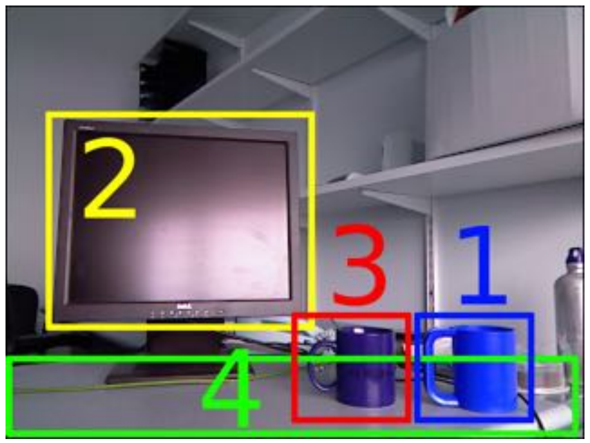
chair (1,brown, image 4, X,Y,Z)

chair (2,brown, image 4, X,Y,Z)

chair (1,brown, image 5, X,Y,Z)

⋮

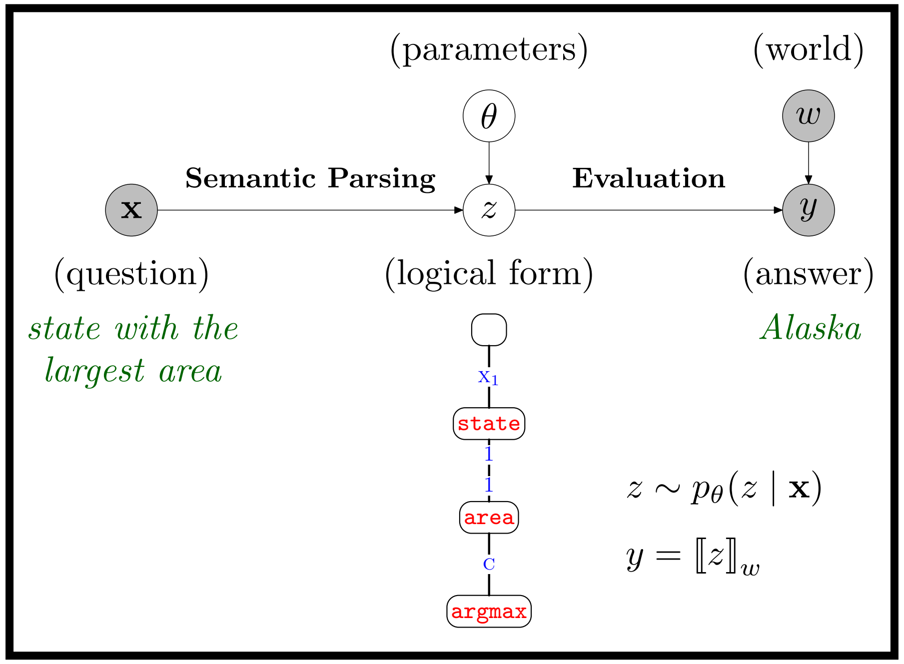
Results

Environment d	Language z and predicted logical form ℓ	Predicted grounding	True grounding
	monitor to the left of the mugs $\lambda x.\exists y.\mathtt{monitor}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$	$\{(2,1),(2,3)\}$	$\{(2,1),(2,3)\}$
	mug to the left of the other mug $\lambda x.\exists y.\mathtt{mug}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$	$\{(3,1)\}$	$\{(3,1)\}$
	objects on the table $\lambda x.\exists y.\mathtt{object}(x) \wedge \mathtt{on\text{-}rel}(x,y) \wedge \mathtt{table}(y)$	$\{(1,4),(2,4)$ $(3,4)\}$	$\{(1,4),(2,4),$ $(3,4)\}$
	two blue cups are placed near to the computer screen $\lambda x.\mathtt{blue}(x) \wedge \mathtt{cup}(x) \wedge \mathtt{comp.}(x) \wedge \mathtt{screen}(x)$	$\{(1)\}$	$\{(1,2),(3,2)\}$

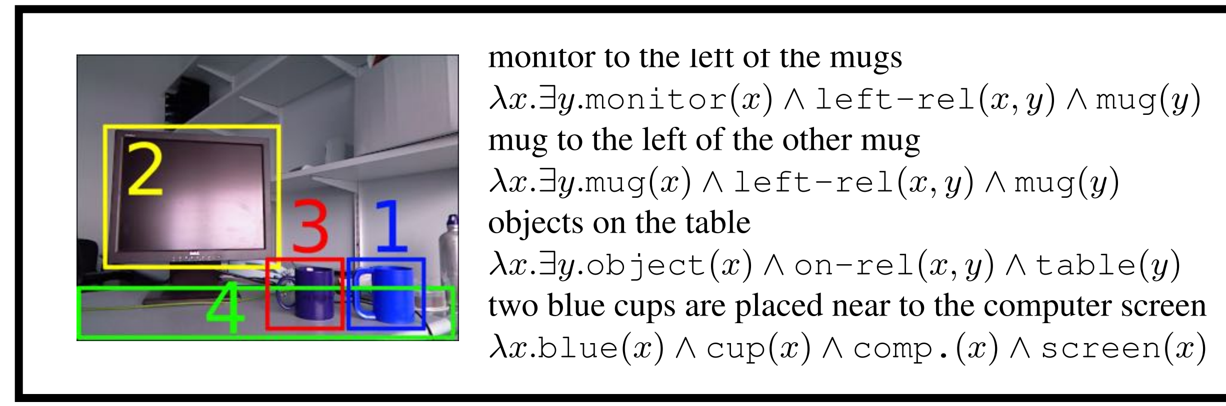
Denotation γ	0 rel.	1 rel.	other	total
LSP-CAT	**0.94**	0.45	**0.20**	0.51
LSP-F	0.89	**0.81**	**0.20**	**0.70**
LSP-W	0.89	0.77	0.16	0.67
Grounding g	0 rel.	1 rel.	other	total
LSP-CAT	**0.94**	0.37	0.00	0.42
LSP-F	0.89	**0.80**	0.00	**0.65**
LSP-W	0.89	0.70	0.00	0.59
% of data	23	56	21	100

(a) Results on the SCENE data set.

Roadmap



Learning Dependency-Based Compositional Semantics
(P. Liang et. al. ACL 2011)



Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World.
(J. Krishnamurthy et. al. TACL 2013)

Some ideas

?

Current limitations

- Language
  - At most 1 relation
  - Doesn't model more complex phenomena (negations, superlatives, …)
- Vision
  - Dataset is restricted
  - No uncertainty

- A computer system is on the table
- There are items on the desk
- There are two cups on the table
- The computer is off

Current limitations

- Language
  - At most 1 relation
  - Doesn't model more complex phenomena (negations, superlatives, …)
- Vision
  - Dataset is restricted
  - No uncertainty

Our suggestions

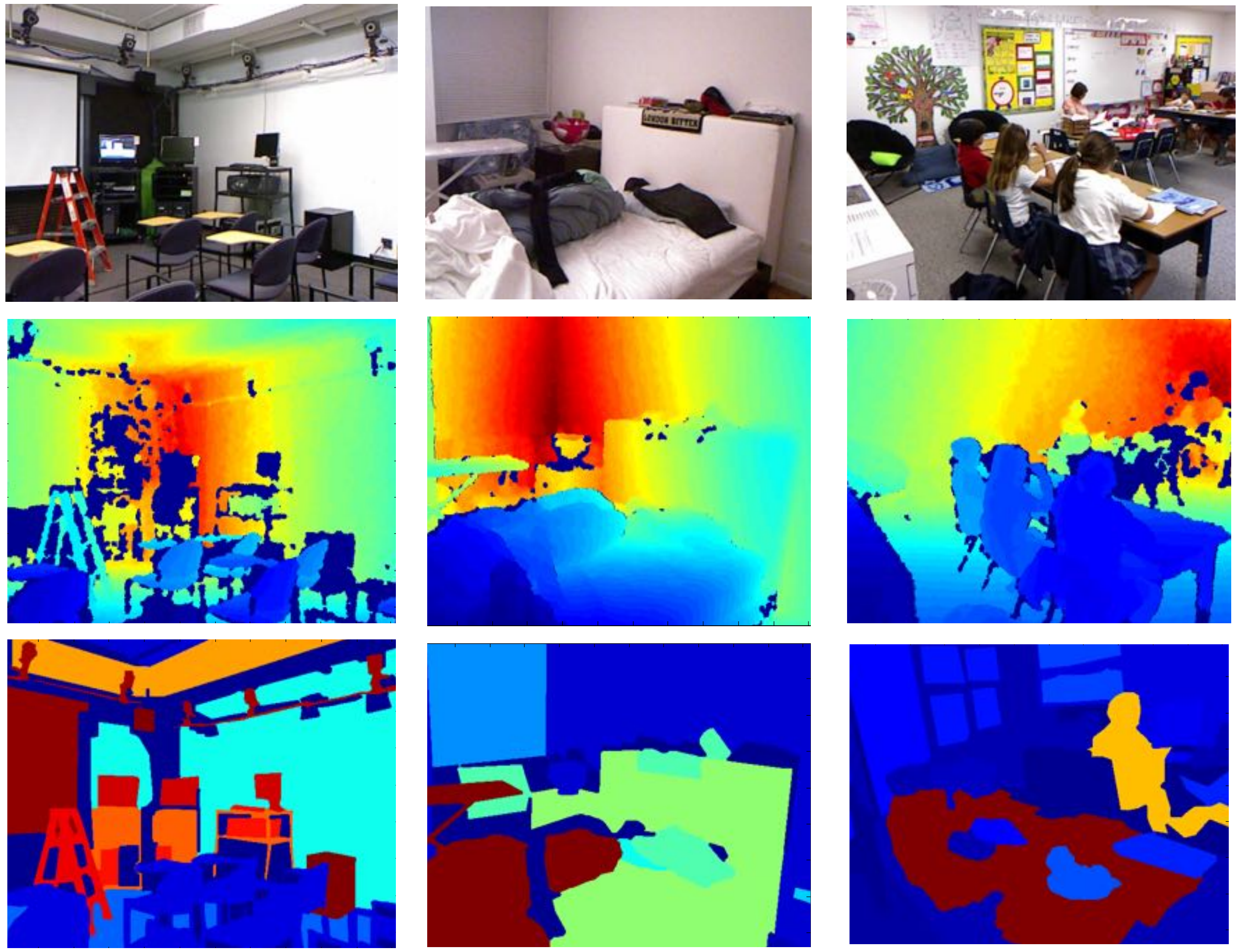
- Language
  - At most 1 relation
  - Doesn't model more complex phenomena (negations, superlatives, …)
- Vision
  - Dataset is restricted
  - No uncertainty

- A computer system is on the table
- There are items on the desk
- There are two cups on the table
- The computer is off

- What is the object in front of the photocopying machine attached to the wall?
- What is the object that is placed on the middle rack of the stand that is placed closed to the wall?
- What is time showing on the clock?

Our suggestions

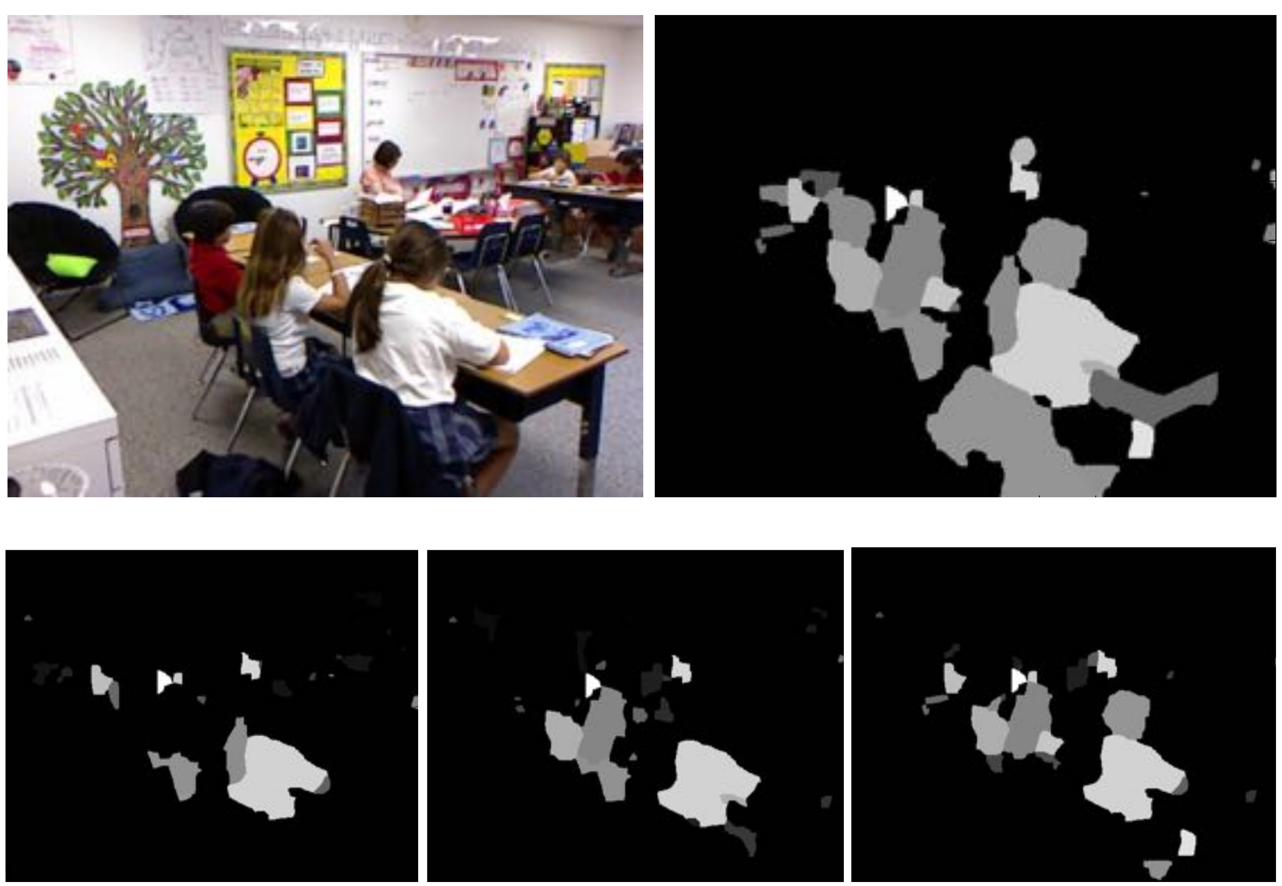
- Language
  - At most 1 relation
  - Doesn't model more complex phenomena (negations, superlatives, …)
- Vision
  - Dataset is restricted
  - No uncertainty



- Indoor Segmentation and Support Inference from RGBD Images (Silberman et. al. ECCV'12)
- Perceptual organization and recognition of indoor scenes from rgb-d images (Gupta et. al. CVPR'13)

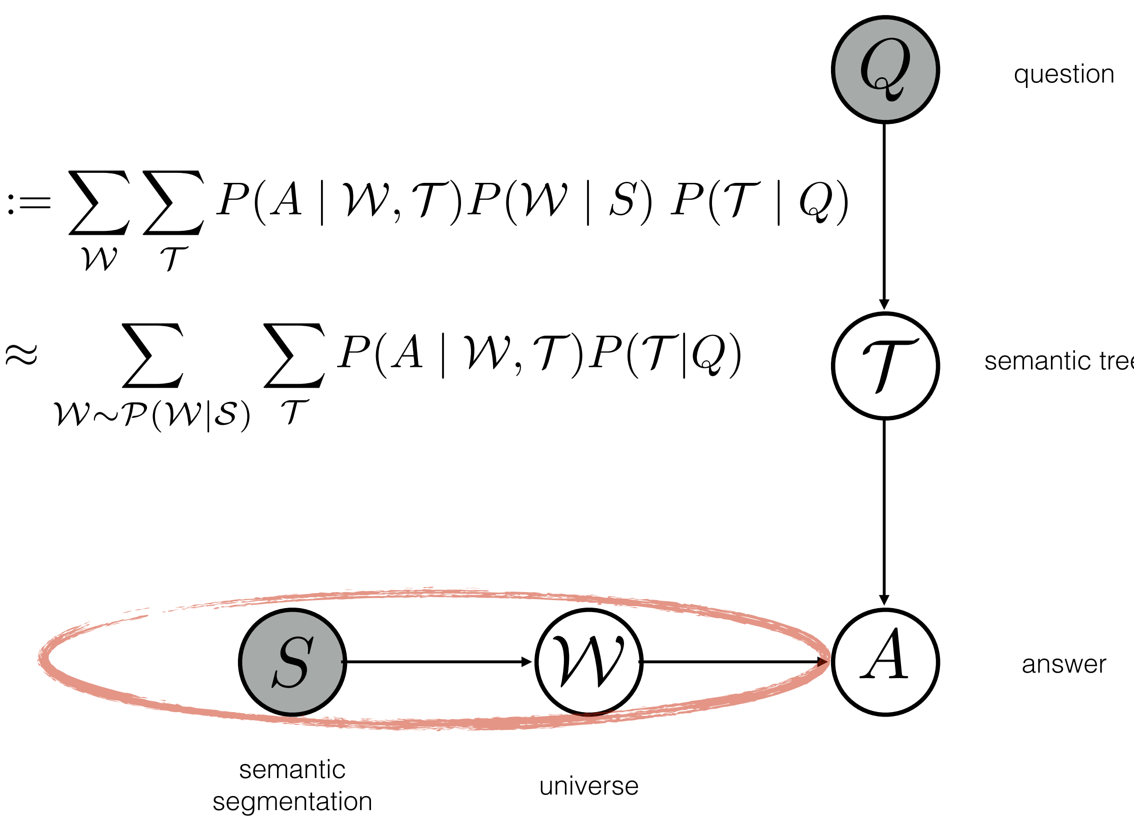
Our suggestions

- Language
  - At most 1 relation
  - Doesn't model more complex phenomena (negations, superlatives, …)
- Vision
  - Dataset is restricted
  - No uncertainty

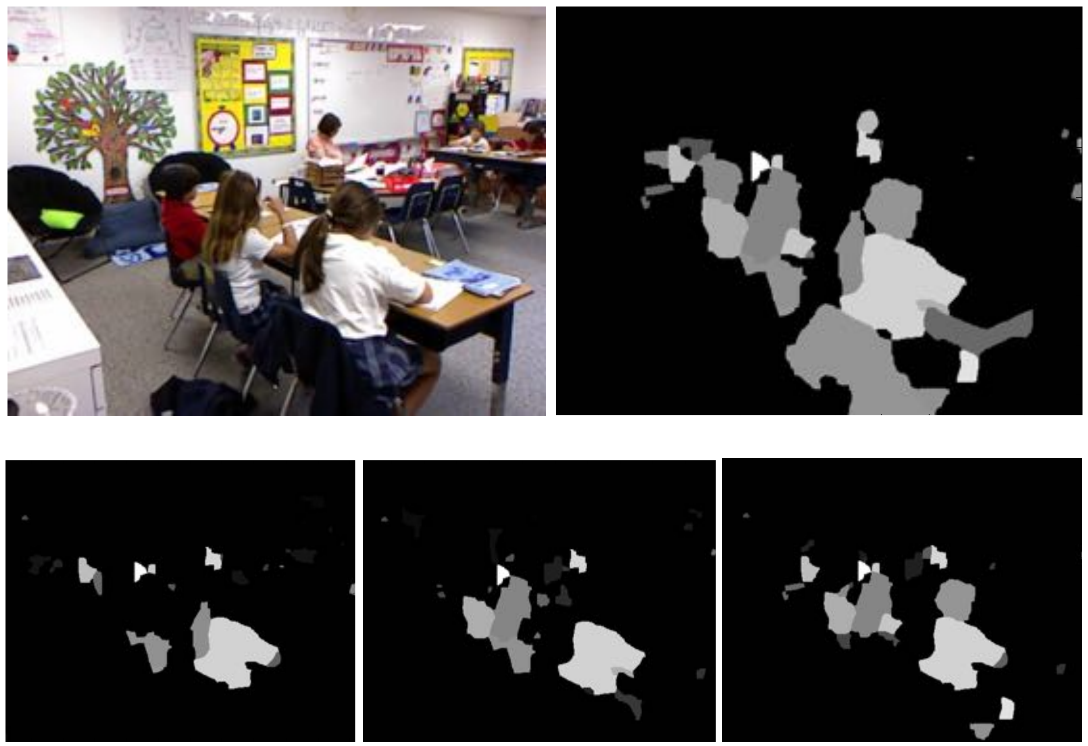


$$P(A \mid Q,S) := \sum_{\mathcal{W}} \sum_{\mathcal{T}} P(A \mid \mathcal{W},\mathcal{T}) P(\mathcal{W} \mid S)\ P(\mathcal{T} \mid Q)$$

$$P(A \mid Q,S) \approx \sum_{\mathcal{W}\sim\mathcal{P}(\mathcal{W}|\mathcal{S})} \sum_{\mathcal{T}} P(A \mid \mathcal{W},\mathcal{T}) P(\mathcal{T}|Q)$$

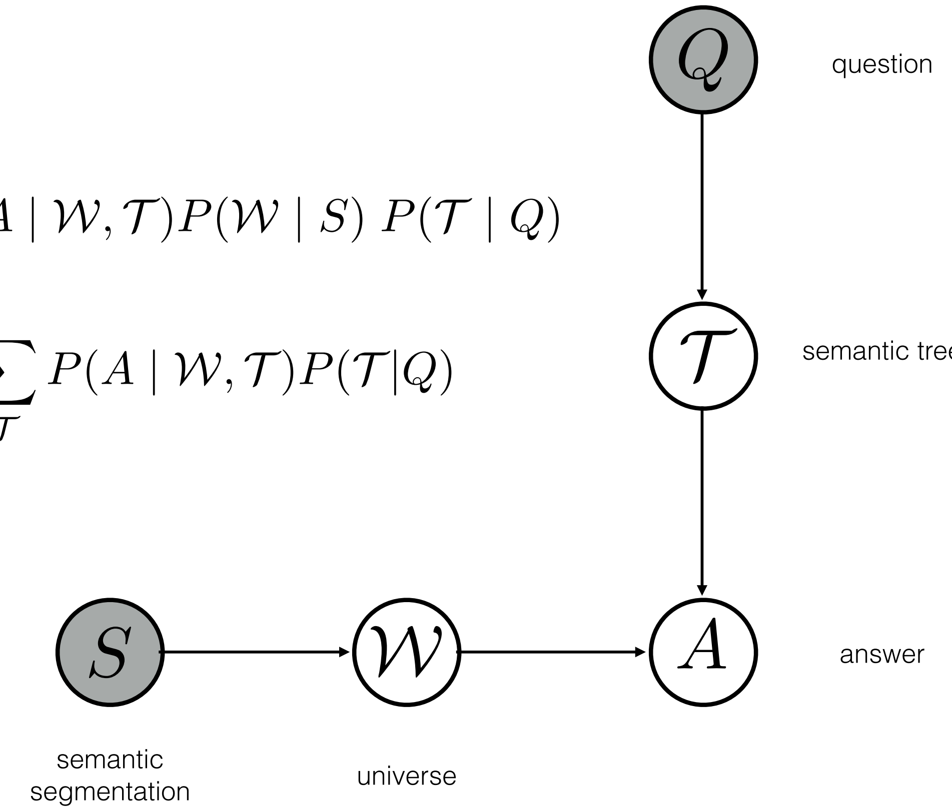


Results



$$P(A \mid Q, S) := \sum_{\mathcal{W}} \sum_{\mathcal{T}} P(A \mid \mathcal{W}, \mathcal{T}) P(\mathcal{W} \mid S)\ P(\mathcal{T} \mid Q)$$

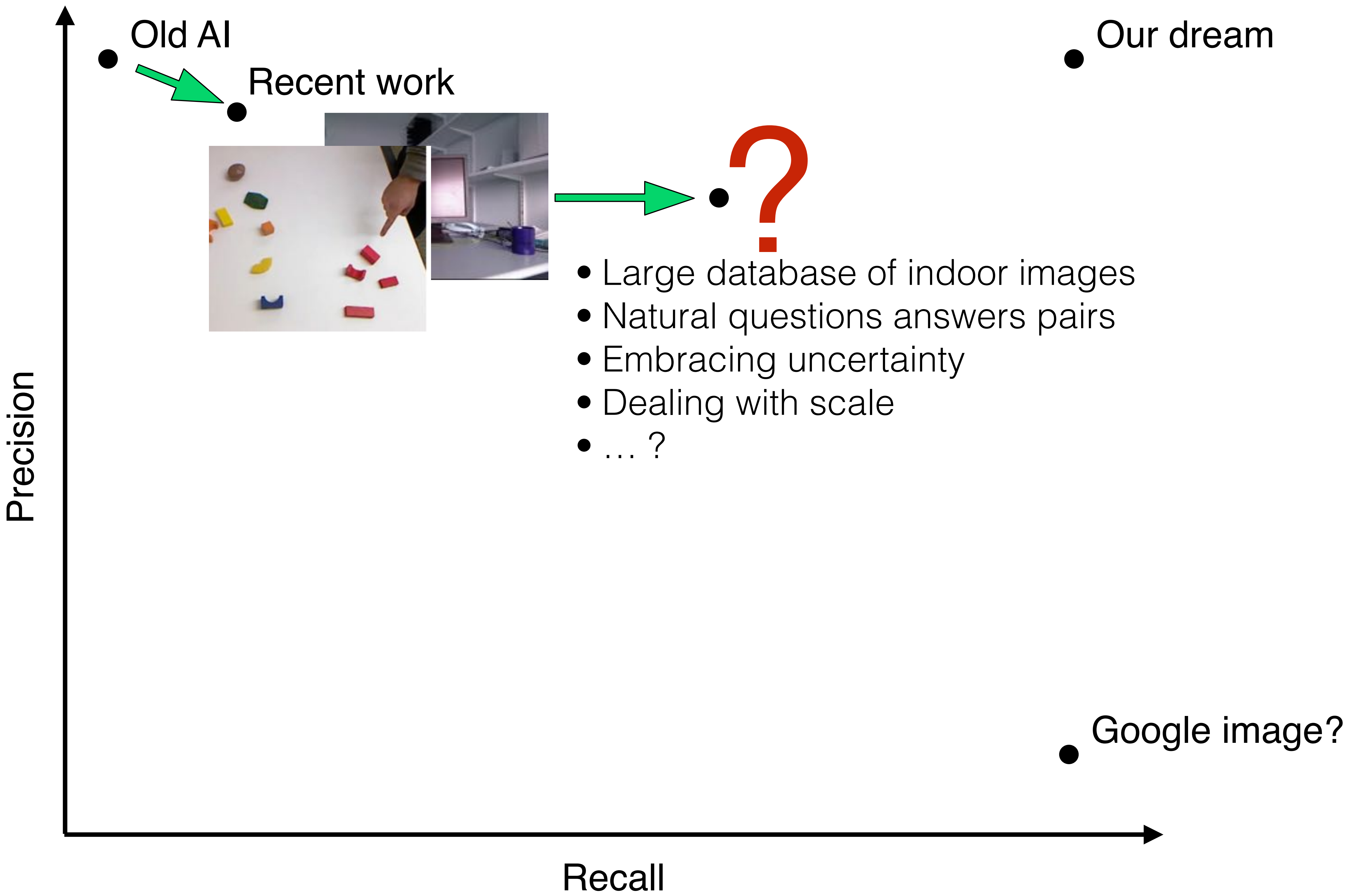
$$P(A \mid Q, S) \approx \sum_{\mathcal{W} \sim \mathcal{P}(\mathcal{W} \mid \mathcal{S})} \sum_{\mathcal{T}} P(A \mid \mathcal{W}, \mathcal{T}) P(\mathcal{T} | Q)$$



Description	Examples
Individual images	
counting	How many cabinets are in image1?
counting and colors	How many gray cabinets are in image1?
room type	Which type of the room is depicted in image1?
superlatives	What is the largest object in image1?
Set of images	
counting and colors	How many black bags?
negations type 1	Which images do not have sofa?
negations type 2	Which images are not bedroom?

Experiments	Accuracy
Perfect detections	56%
One universe	11.25%
Multiuniverse	13.75%

Two dimensions of question answering challenge



max planck institut
informatik

Visual Turing Test: ongoing challenge

Mateusz Malinowski

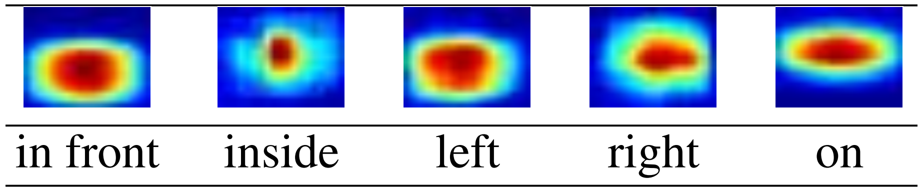
Visual question answering challenge

Ask about the content of the image

- How many sofas? ⟶ 3
- Where is the lamp? ⟶ on the table, close to tv
- What is behind the largest table? ⟶ tv
- What is the color of the walls? ⟶ purple

The task involves

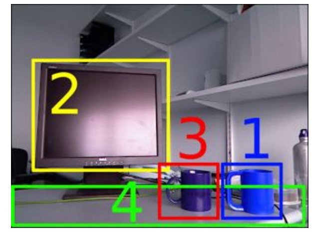
Object detection



Spatial reasoning

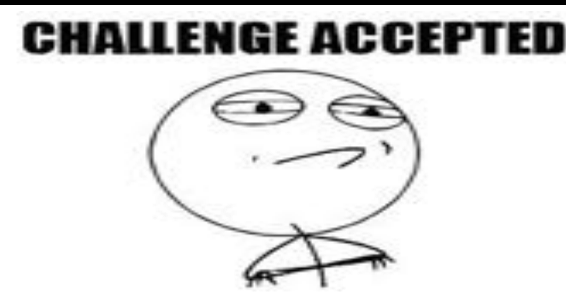
Natural language understanding

Outline



monitor to the left of the mugs
$\lambda x.\exists y.\mathtt{monitor}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
mug to the left of the other mug
$\lambda x.\exists y.\mathtt{mug}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
objects on the table
$\lambda x.\exists y.\mathtt{object}(x) \wedge \mathtt{on\text{-}rel}(x,y) \wedge \mathtt{table}(y)$
two blue cups are placed near to the computer screen
$\lambda x.\mathtt{blue}(x) \wedge \mathtt{cup}(x) \wedge \mathtt{comp.}(x) \wedge \mathtt{screen}(x)$

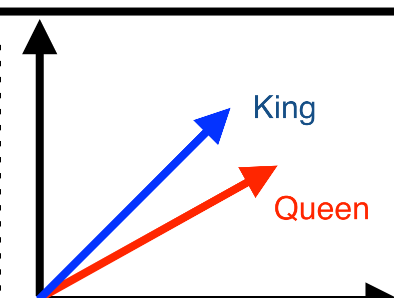
State-of-the-art



Challenges

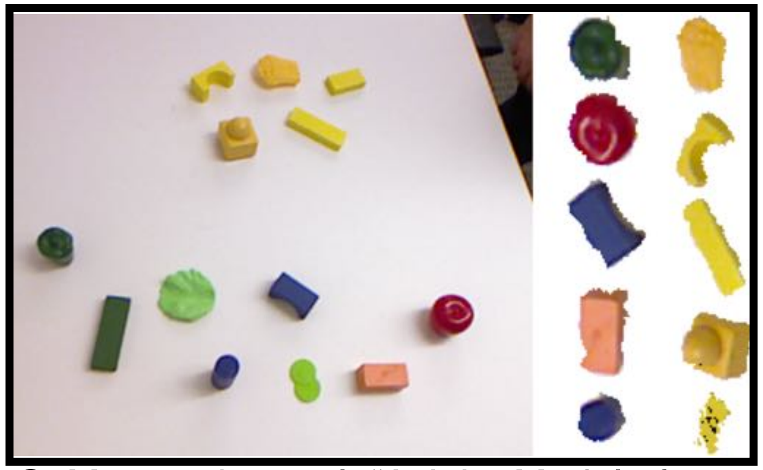
Natural Language Understanding

monitor to the left of the mugs
$\lambda x.\exists y.\mathtt{monitor}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
mug to the left of the other mug
$\lambda x.\exists y.\mathtt{mug}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
objects on the table
$\lambda x.\exists y.\mathtt{object}(x) \wedge \mathtt{on\text{-}rel}(x,y) \wedge \mathtt{table}(y)$
two blue cups are placed near to the computer screen
$\lambda x.\mathtt{blue}(x) \wedge \mathtt{cup}(x) \wedge \mathtt{comp.}(x) \wedge \mathtt{screen}(x)$

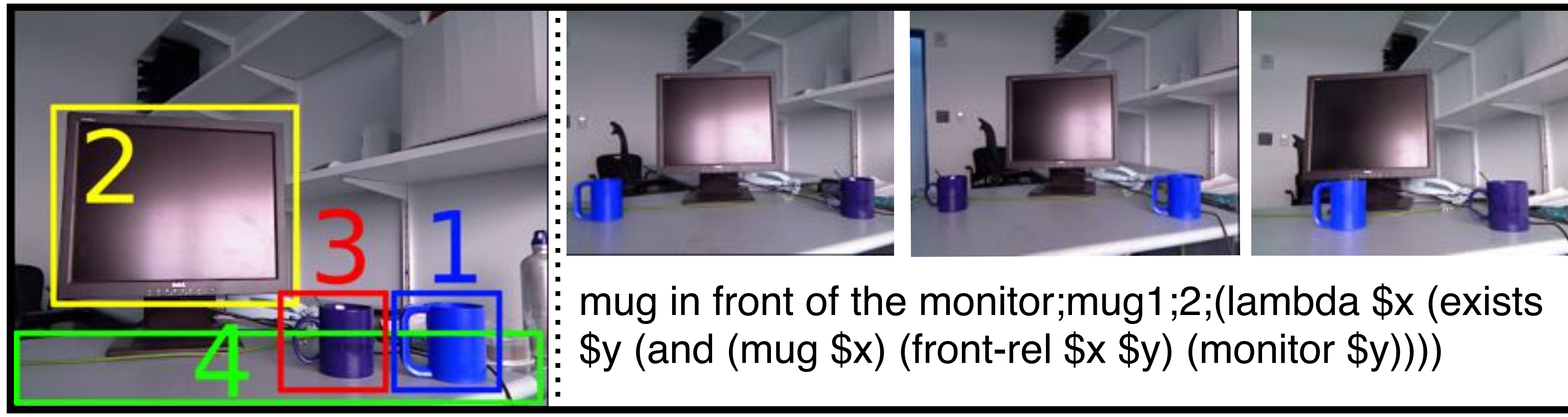


Two extremes on language understanding

From language grounding to question answering



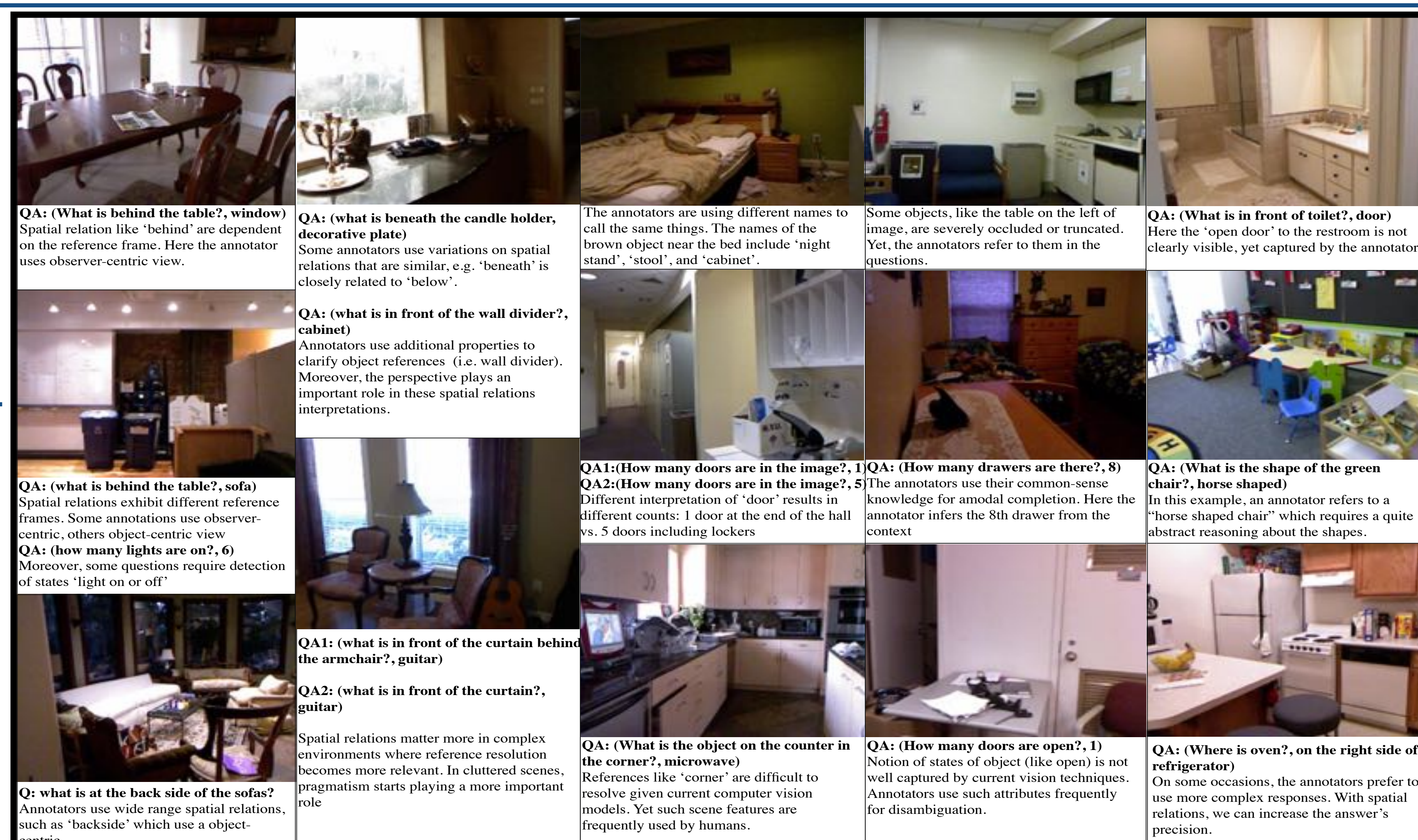
C. Matuszek, et. al. “A Joint Model of Language and Perception Grounded Attribute Learning” ICML 2012



J. Krishnamurthy, et. al. “Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World” TACL 2013

- More real-world images
- More categories
- More questions, answers
- More question types
- No logical forms

- Different than grounding
- ‘Social consensus’, not ‘connecting to the physical world’
- Latent motivations of the questioner



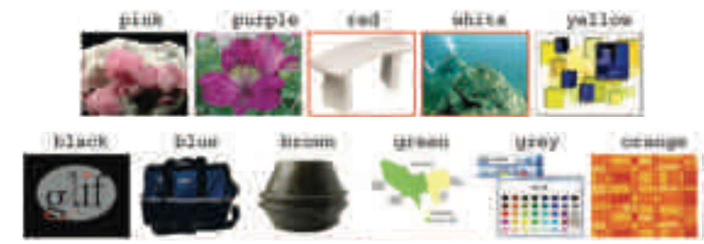
N. Silberman, et. al. NYU Depth Dataset V2 ECCV 2012

Briefly about the approach

P. Liang, et. al. “Learning dependency-based compositional semantics” ACL 2011

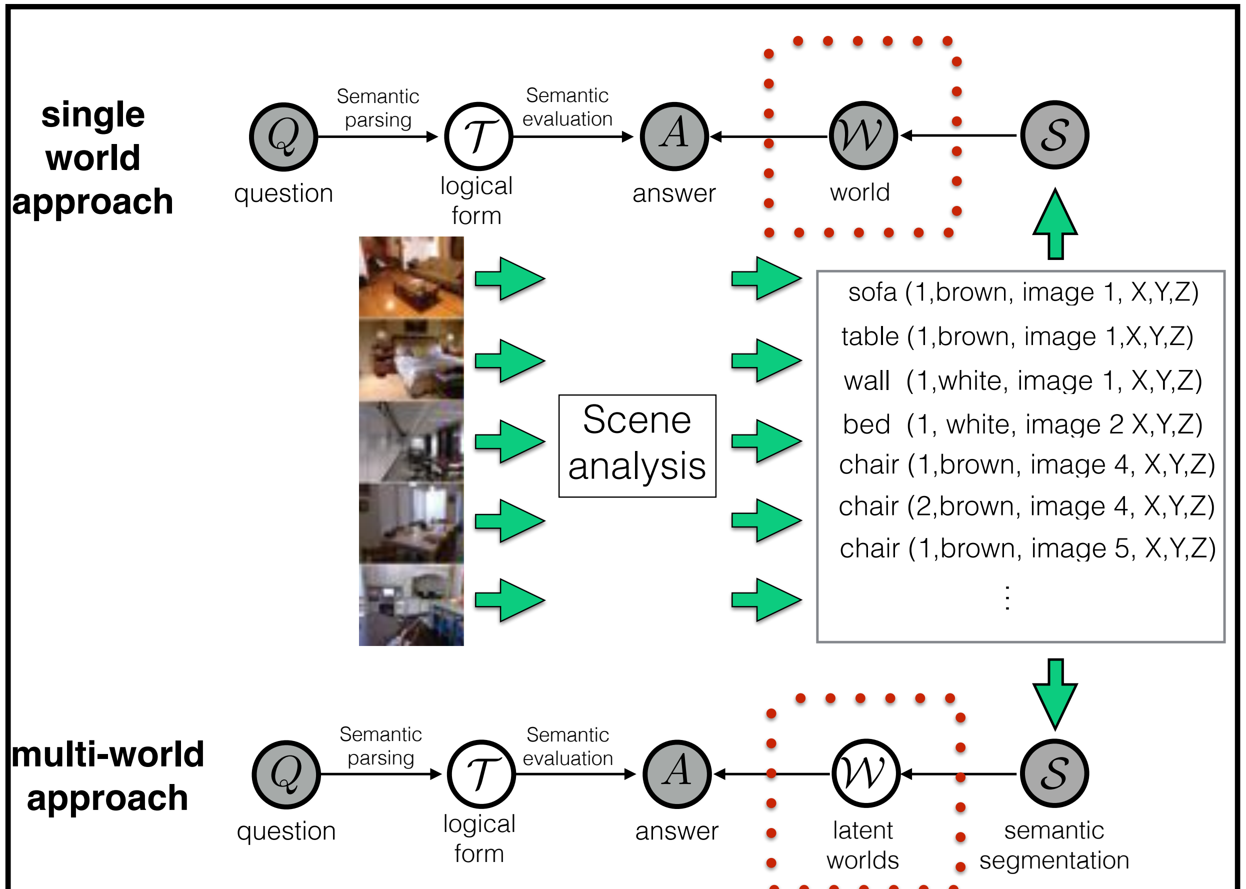
+

S. Gupta, et. al. “Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images” CVPR 2013



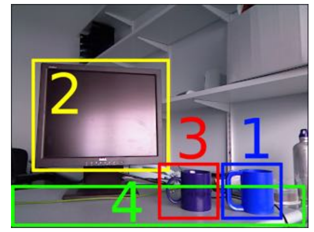
J. Weijer, et. al. “Learning Color Names for Real World Applications” TIP 2009

Scene analysis



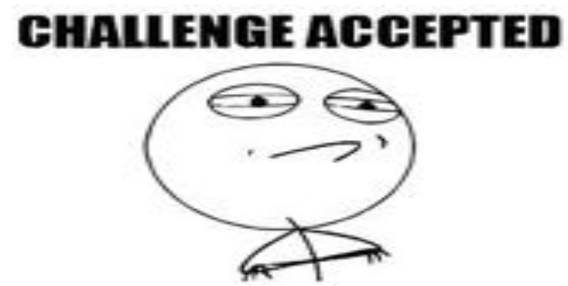
$$P(A \mid Q, S) = \sum_{\mathcal{W}} \sum_{\mathcal{T}} P(A \mid \mathcal{W}, \mathcal{T}) P(\mathcal{W} \mid S)\ P(\mathcal{T} \mid Q)$$

Outline



monitor to the left of the mugs
$\lambda x.\exists y.\mathtt{monitor}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
mug to the left of the other mug
$\lambda x.\exists y.\mathtt{mug}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
objects on the table
$\lambda x.\exists y.\mathtt{object}(x) \wedge \mathtt{on\text{-}rel}(x,y) \wedge \mathtt{table}(y)$
two blue cups are placed near to the computer screen
$\lambda x.\mathtt{blue}(x) \wedge \mathtt{cup}(x) \wedge \mathtt{comp.}(x) \wedge \mathtt{screen}(x)$

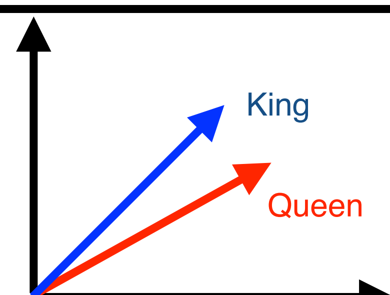
State-of-the-art



Challenges

Natural Language Understanding

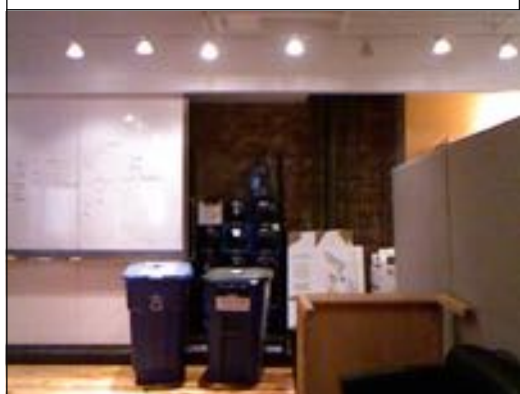
monitor to the left of the mugs
$\lambda x.\exists y.\mathtt{monitor}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
mug to the left of the other mug
$\lambda x.\exists y.\mathtt{mug}(x) \wedge \mathtt{left\text{-}rel}(x,y) \wedge \mathtt{mug}(y)$
objects on the table
$\lambda x.\exists y.\mathtt{object}(x) \wedge \mathtt{on\text{-}rel}(x,y) \wedge \mathtt{table}(y)$
two blue cups are placed near to the computer screen
$\lambda x.\mathtt{blue}(x) \wedge \mathtt{cup}(x) \wedge \mathtt{comp.}(x) \wedge \mathtt{screen}(x)$



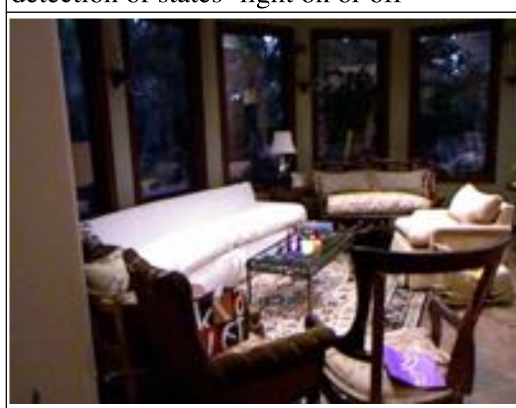
Two extremes on language understanding

Challenges

QA: (What is behind the table?, window)
Spatial relation like 'behind' are dependent on the reference frame. Here the annotator uses observer-centric view.



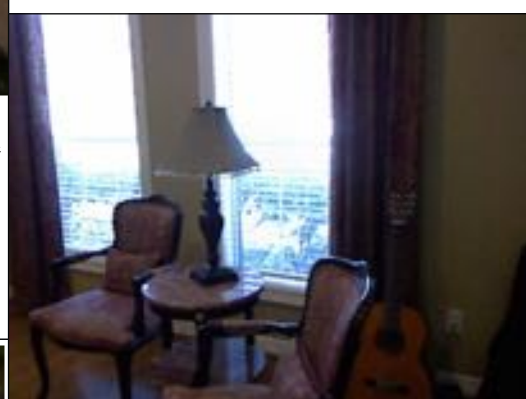
QA: (what is behind the table?, sofa)
Spatial relations exhibit different reference frames. Some annotations use observer-centric, others object-centric view
QA: (how many lights are on?, 6)
Moreover, some questions require detection of states 'light on or off'



Q: what is at the back side of the sofas?
Annotators use wide range spatial relations, such as 'backside' which is object-centric.

QA: (what is beneath the candle holder, decorative plate)
Some annotators use variations on spatial relations that are similar, e.g. 'beneath' is closely related to 'below'.

QA: (what is in front of the wall divider?, cabinet)
Annotators use additional properties to clarify object references (i.e. wall divider). Moreover, the perspective plays an important role in these spatial relations interpretations.



QA1: (what is in front of the curtain behind the armchair?, guitar)

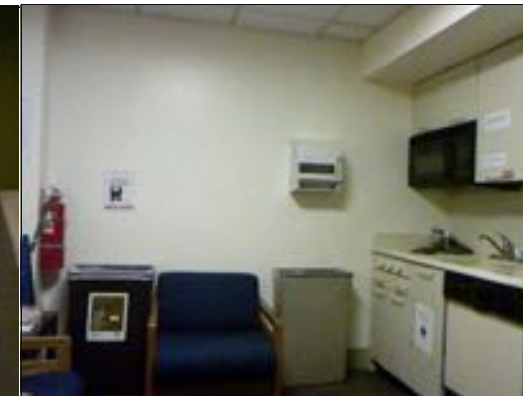
QA2: (what is in front of the curtain?, guitar)

Spatial relations matter more in complex environments where reference resolution becomes more relevant. In cluttered scenes, pragmatism starts playing a more important role

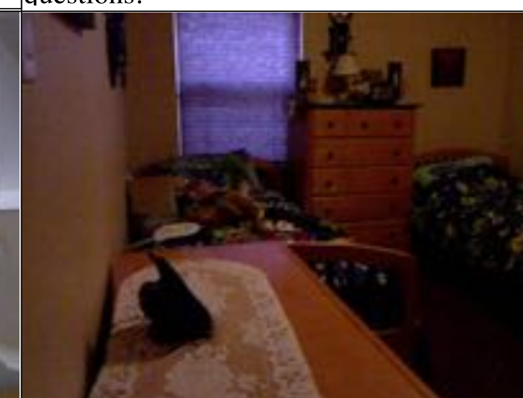
The annotators are using different names to call the same things. The names of the brown object near the bed include 'night stand', 'stool', and 'cabinet'.

QA1:(How many doors are in the image?, 1)
QA2:(How many doors are in the image?, 5)
Different interpretation of 'door' results in different counts: 1 door at the end of the hall vs. 5 doors including lockers

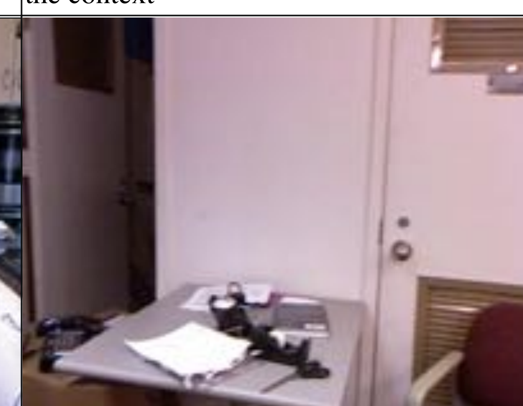
QA: (What is the object on the counter in the corner?, microwave)
References like 'corner' are difficult to resolve given current computer vision models. Yet such scene features are frequently used by humans.



Some objects, like the table on the left of image, are severely occluded or truncated. Yet, the annotators refer to them in the questions.



QA: (How many drawers are there?, 8)
The annotators use their common-sense knowledge for amodal completion. Here the annotator infers the 8th drawer from the context



QA: (How many doors are open?, 1)
Notion of states of object (like open) is not well captured by current vision techniques. Annotators use such attributes frequently for disambiguation.

QA: (What is in front of toilet?, door)
Here the 'open door' to the restroom is not clearly visible, yet captured by the annotator.

QA: (What is above the desk in front of the scissors?, hole puncher)
It is difficult to find the scissors solely with the appearance-based methods.

QA: (Where is oven?, on the right side of refrigerator)
On some occasions, the annotators prefer to use more complex responses. With spatial relations, we can increase the answer's precision.

Other challenges

- Detectors for more categories
  - Currently 37 categories, but we need about 900
- Metric to benchmark methods
  - Semantic boundaries between the categories becomes unclear
    - carton ~ box
    - cup ~ cup of coffee
  - This suggests a metric that is built on some ontologies
    - Wu-Palmer similarity on the WordNet taxonomy
    - Problems with WordNet: ‘garbage bin’ doesn’t exist
  - Takes into account ‘social consensus’
    - Possible different answers
    - Ongoing work
  - Metric: $\mathrm{WUPS}(A,T) = \frac{1}{N}\sum_{i=1}^{N}\min\{\prod_{a\in A^i}\max_{t\in T^i}\mathrm{WUP}(a,t),\ \prod_{t\in T^i}\max_{a\in A^i}\mathrm{WUP}(a,t)\}\cdot 100$
- Problems with the semantic parser

Results

Description	Template
counting	How many {object} are in {image_id}?
counting and colors	How many {color} {object} are in {image_id}?
room type	Which type of the room is depicted in {image_id}?
superlatives	What is the largest {object} in {image_id}?
counting and colors	How many {color} {object}?
negations type 1	Which images do not have {object}?
negations type 2	Which images are not {room_type}?
negations type 3	Which images have {object} but do not have a {object}?

synthetic question-answer pairs (SynthQA)

Segmentation	World(s)	# classes	Accuracy
HumanSeg	Single with Neg. 3	37	56.0%
HumanSeg	Single	37	59.5%
AutoSeg	Single	37	11.25%
AutoSeg	Multi	37	13.75%

Human question-answer pairs (HumanQA)

Segmentation	World(s)	#classes	Accuracy	WUPS at 0.9	WUPS at 0
HumanSeg	Single	894	7.86%	11.86%	38.79%
HumanSeg	Single	37	12.47%	16.49%	50.28%
AutoSeg	Single	37	9.69%	14.73%	48.57%
AutoSeg	Multi	37	12.73%	18.10%	51.47%
Human Baseline		894	50.20%	50.82%	67.27%
Human Baseline		37	60.27%	61.04%	78.96%

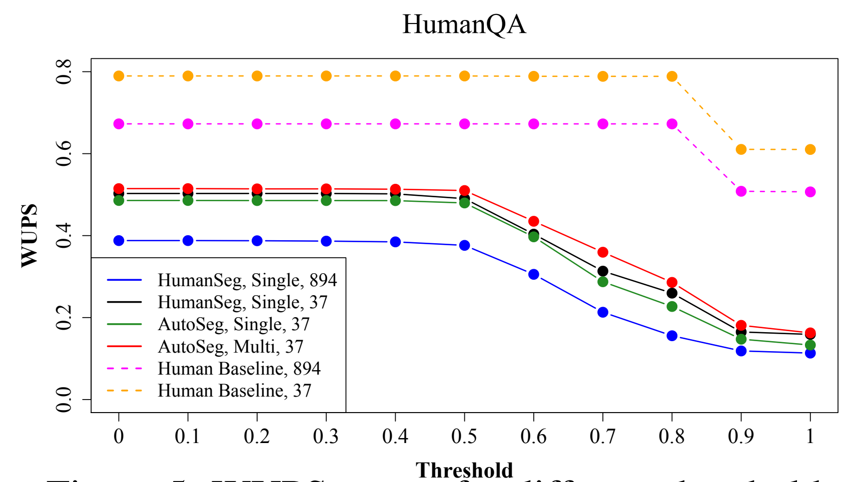


Figure 5: WUPS scores for different thresholds.

Q: How many red chairs are there?
H: ()
M: 6
C: blinds

Q: How many chairs are at the table?
H: wall
M: 4
C: chair

Q: What is on the right side of cabinet?
H: picture
M: bed
C: bed

Q: What is on the wall?
H: mirror
M: bed
C: picture

Q: What is the object on the chair?
H: pillow
M: floor, wall
C: wall

Q: What is on the right side of the table?
H: chair
M: window, floor, wall
C: floor

Q: What is behind the television?
H: lamp
M: brown, pink, purple
C: picture

Q: What is in front of television?
H: pillow
M: chair
C: picture

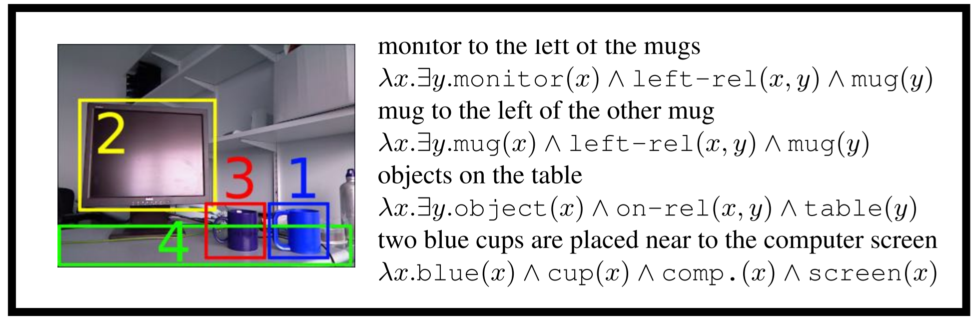
Q: What color is the bed?
H: black, blue, …

Q: What color is the pillow?
H: blue

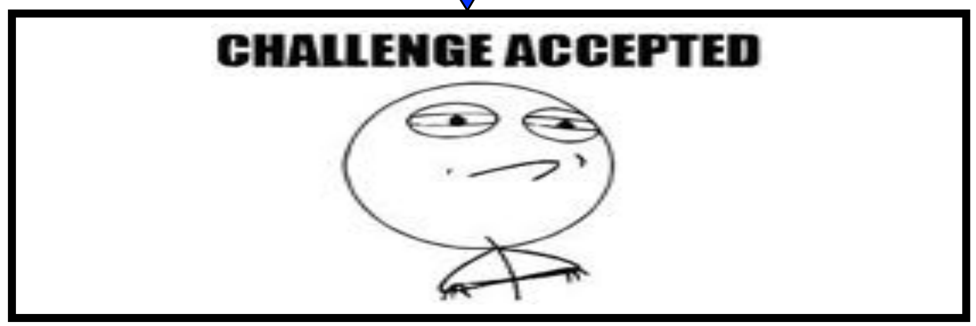
Q: What color is the bed?
H: blue

Q: What color is the pillow?
H: red

Outline

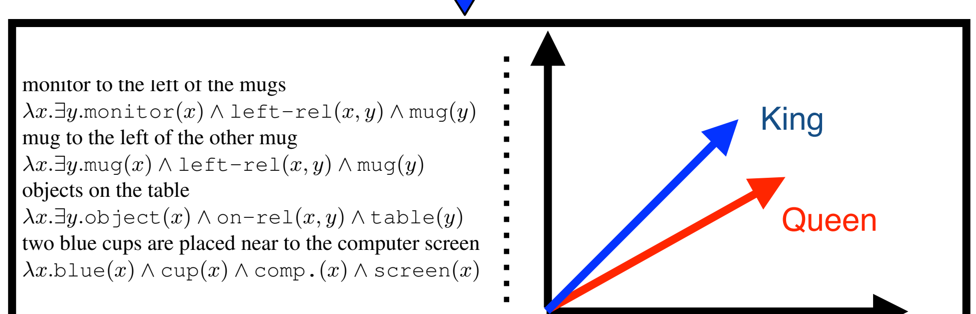


State-of-the-art



Challenges

Natural Language Understanding

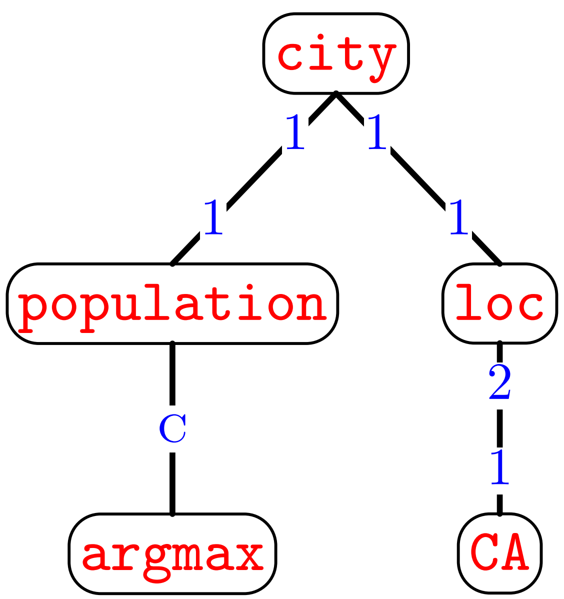


Two extremes on language understanding

Words to Predicates (Lexical Semantics)

				city	city			
				state	state			
				river	river			
			argmax	population	population		CA	
What	*is*	*the*	*most*	*populous*	*city*	*in*	*CA*	*?*

most populous city in California



Los Angeles

$$\arg\max_{Pop}\ population(X, Pop), city(X), loc(X, Y), CA(Y)$$

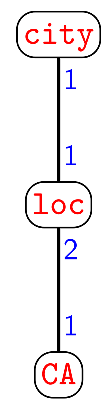
$$city(california, ca, los\ angeles, 2966850)$$
$$state(california, ca, ..., los\ angeles)$$

$$city(cityid(City, St))\ :-\ city\ (\_, St, City, \_)$$
$$population(cityid(City, St), Pop)\ :-\ city(\_, St, City, Pop)$$
$$loc(cityid(City, St), stateid(State))\ :-\ state(State, St, \_, \_, ..., \_, City)$$

Constraint Satisfaction Problem

Basic DCS Trees

DCS tree



Constraints

$c \in \texttt{city}$

$c_1 = \ell_1$

$\ell \in \texttt{loc}$

$\ell_2 = s_1$

$s \in \texttt{CA}$

Database

city

San Francisco
Chicago
Boston
...

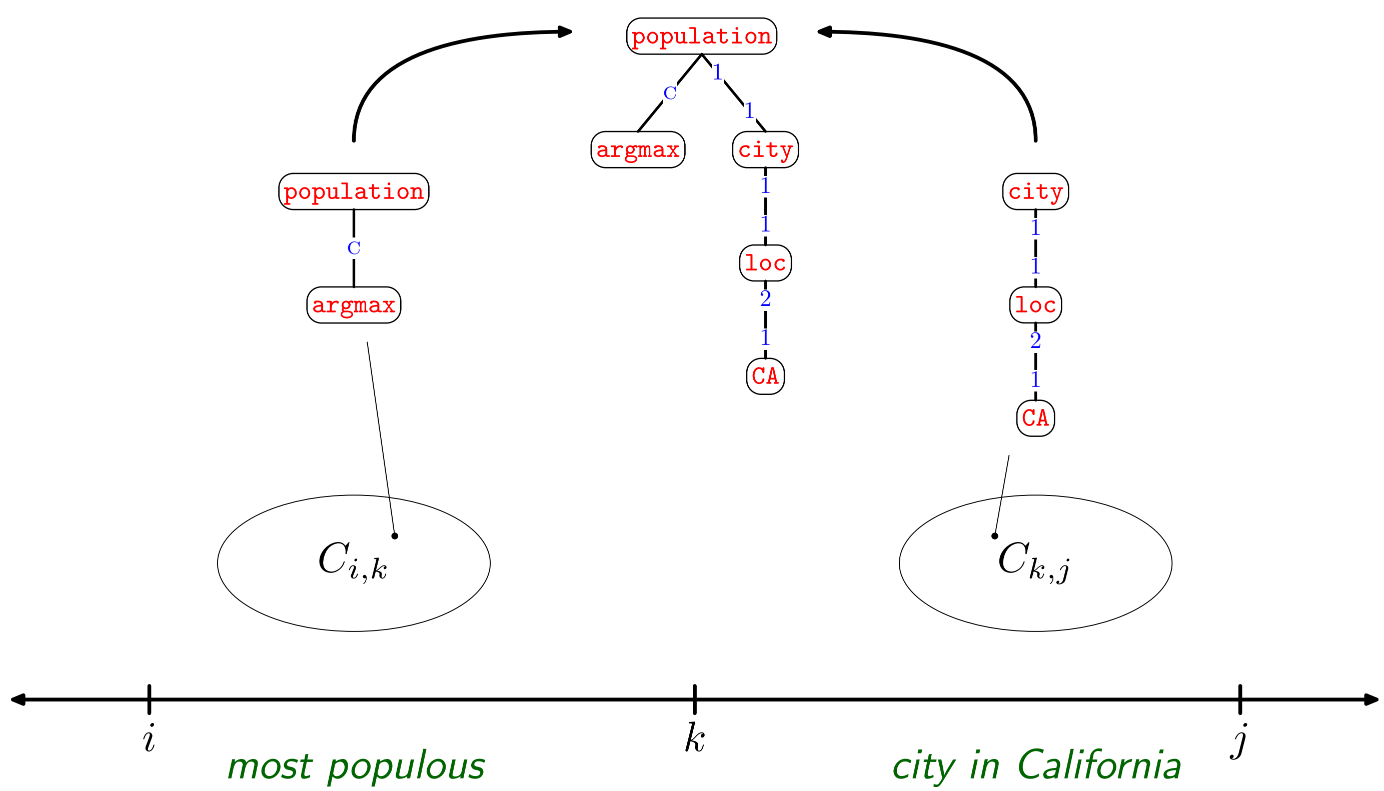
loc

Mount Shasta	California
San Francisco	**California**
Boston	Massachusetts
...	...

CA

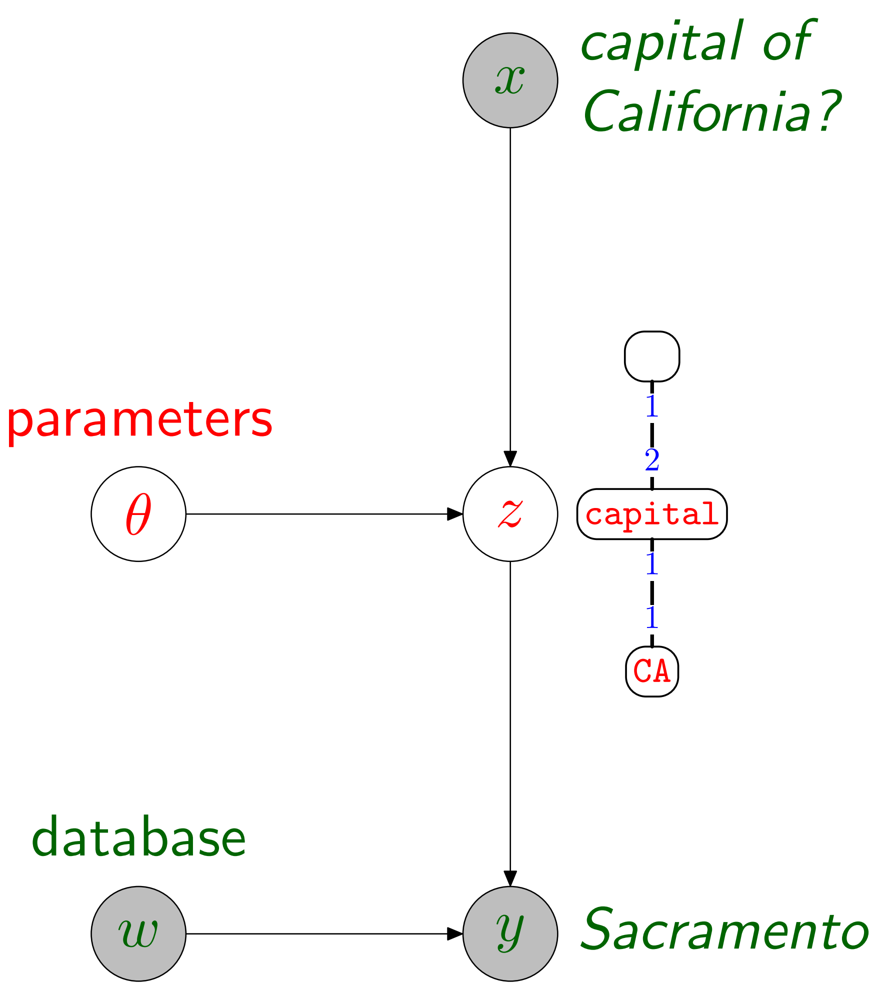
California

Construction Mechanism



Words to Predicates (Lexical Semantics)

What	is	the	most	populous	city	in	CA	?
			argmax	city state river population	city state river population		CA	



Objective

$$\max_{\theta} \sum_{z} p(y \mid z, w)\, p(z \mid x, \theta)$$

Interpretation **Semantic parsing**

Learning

parameters θ

$(0.2, -1.3, \ldots, 0.7)$

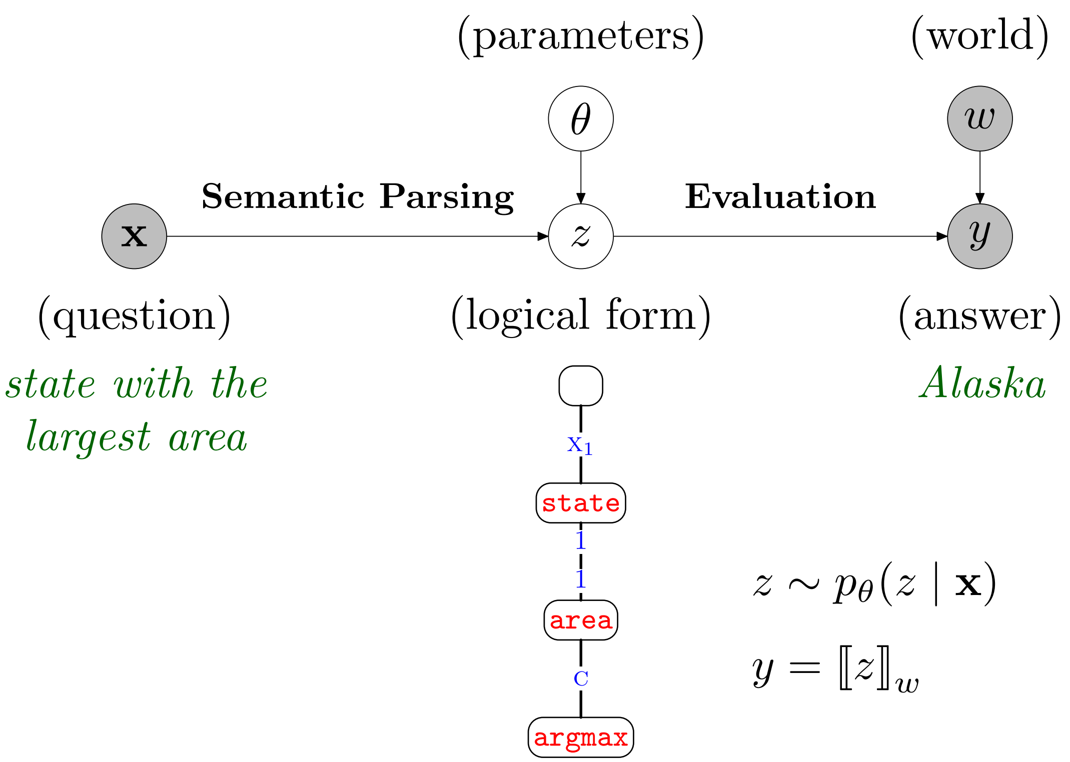
enumerate/score DCS trees →

← numerical optimization (L-BFGS)

k-best list

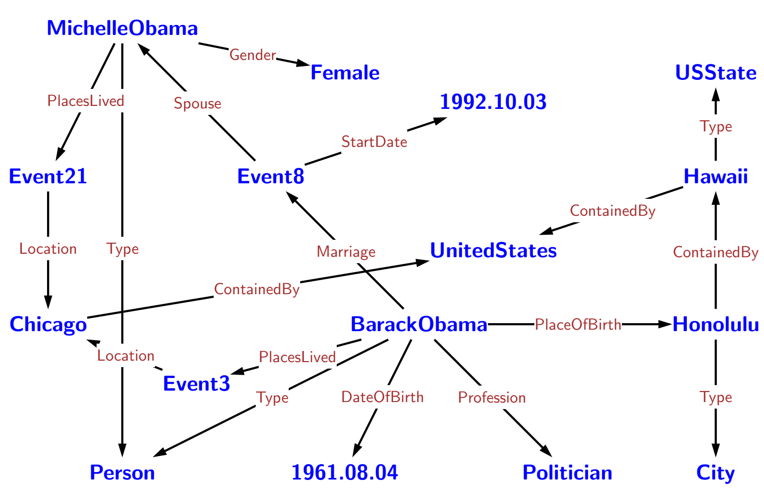
- tree1 ✗
- tree2 ✗
- tree3 ✓
- tree4 ✗
- tree5 ✗

Natural Language Understanding



$z \sim p_\theta(z \mid \mathbf{x})$

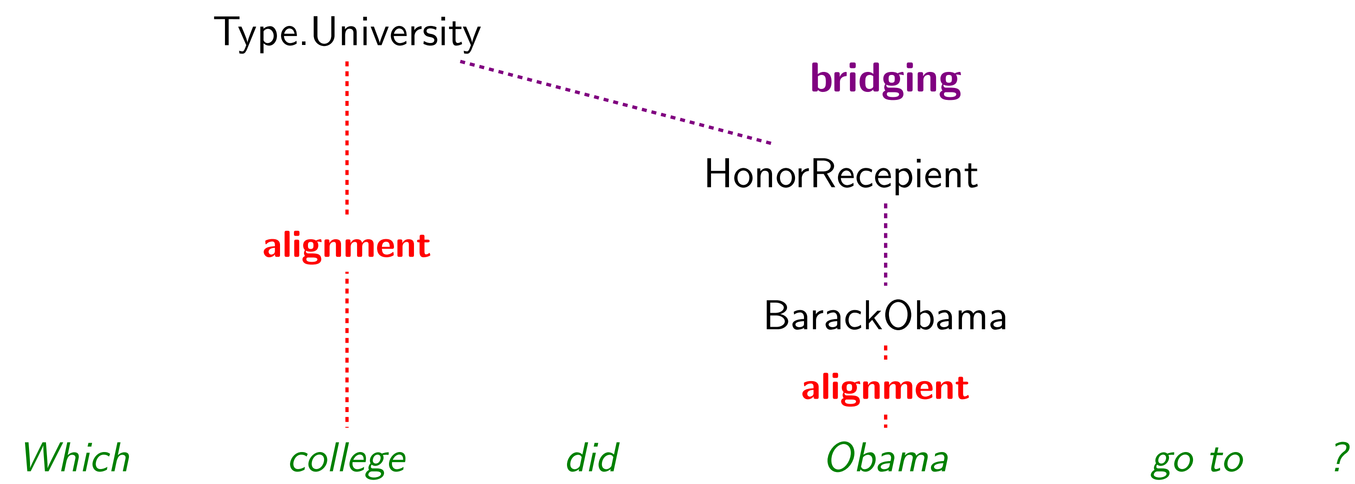
$y = [\![z]\!]_w$



41M entities (nodes)

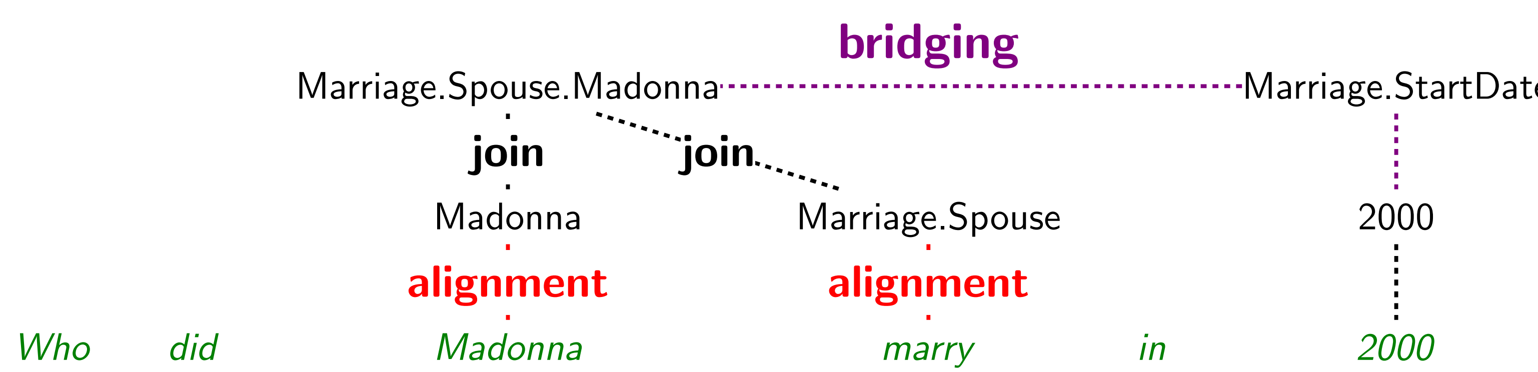
19K properties (edge labels)

596M assertions (edges)



Type.University ⊓ Education.Institution.BarackObama

$z_1 \sqcap b.z_2$ where $z_1 \in t_1,\ z_2 \in t_2,\ b \in (t_1, t_2)$



Marriage.(Spouse.Madonna ⊓ StartDate.2000)

$p_1.(p_2.z' \sqcap b.z)$ where $p_2 \in (t_1, *), z \in t, b \in (t_1, t)$

Results

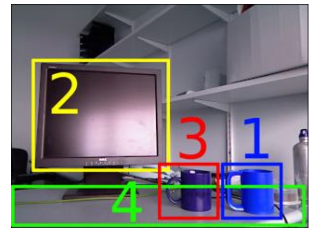
- Examples:
  - How big is Texas?
  - How many states have a city named Springfield?
  - Which rivers run through states bordering New Mexico,?

System	GEO	JOBS
Tang and Mooney (2001)	79.4	79.8
Wong and Mooney (2007)	86.6	–
Zettlemoyer and Collins (2005)	79.3	79.3
Zettlemoyer and Collins (2007)	86.1	–
Kwiatkowski et al. (2010)	88.2	–
Kwiatkowski et al. (2010)	88.9	–
Our system (DCS with L)	88.6	91.4
Our system (DCS with L^+)	**91.1**	**95.0**

System	FREE917	WebQ.
ALIGNMENT	38.0	30.6
BRIDGING	66.9	21.2
ALIGNMENT+BRIDGING	**71.3**	**32.9**

- Web Queries - new large scale dataset with only question, answer pairs
- Google Suggest API is used to build a set of questions
- Questions are sent to AMT workers whose task is to answer on the questions based on the Freebase - in total 5.810 QA pairs
- Examples:
  - What character did Natalie Portman play in Star Wars?
  - What kind of money to take to Bahamas?
  - What did Edward Jenner do for living?

Outline



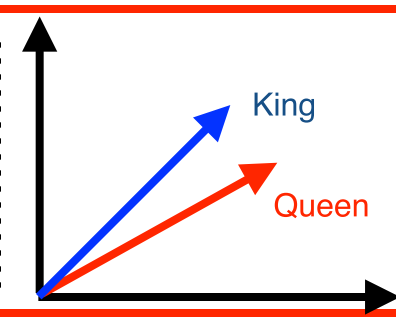
monitor to the left of the mugs
$\lambda x.\exists y.\texttt{monitor}(x) \wedge \texttt{left-rel}(x,y) \wedge \texttt{mug}(y)$
mug to the left of the other mug
$\lambda x.\exists y.\texttt{mug}(x) \wedge \texttt{left-rel}(x,y) \wedge \texttt{mug}(y)$
objects on the table
$\lambda x.\exists y.\texttt{object}(x) \wedge \texttt{on-rel}(x,y) \wedge \texttt{table}(y)$
two blue cups are placed near to the computer screen
$\lambda x.\texttt{blue}(x) \wedge \texttt{cup}(x) \wedge \texttt{comp.}(x) \wedge \texttt{screen}(x)$

State-of-the-art

Challenges

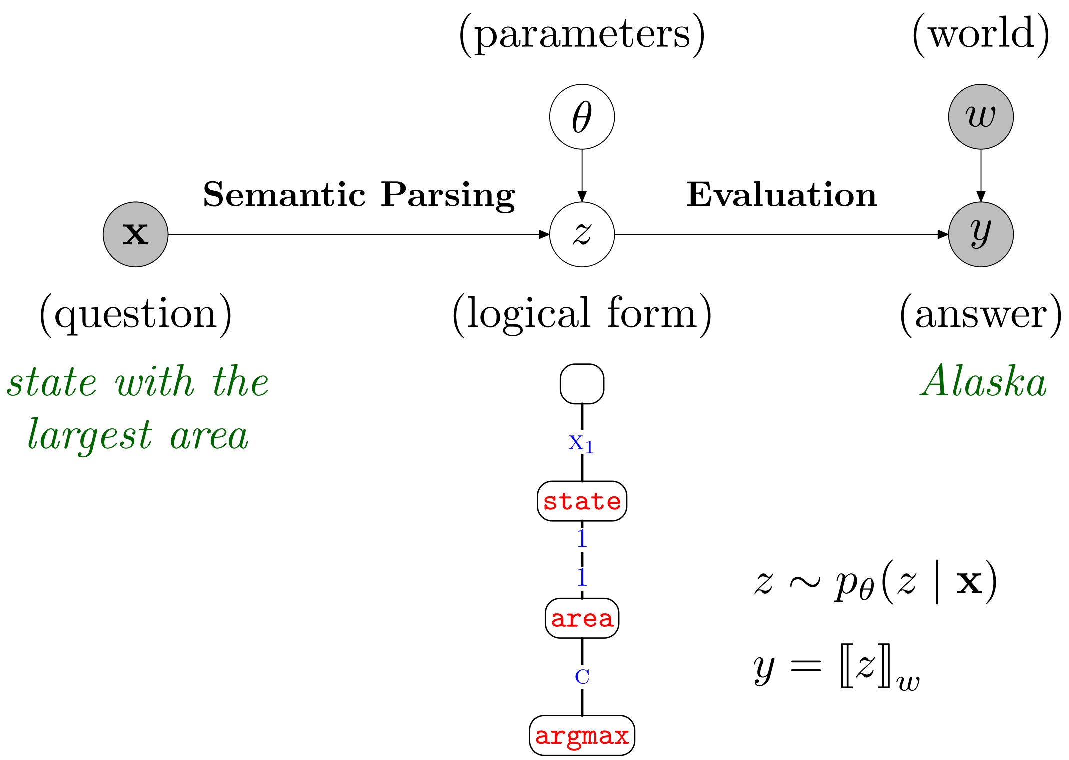
Natural Language Understanding

monitor to the left of the mugs
$\lambda x.\exists y.\texttt{monitor}(x) \wedge \texttt{left-rel}(x,y) \wedge \texttt{mug}(y)$
mug to the left of the other mug
$\lambda x.\exists y.\texttt{mug}(x) \wedge \texttt{left-rel}(x,y) \wedge \texttt{mug}(y)$
objects on the table
$\lambda x.\exists y.\texttt{object}(x) \wedge \texttt{on-rel}(x,y) \wedge \texttt{table}(y)$
two blue cups are placed near to the computer screen
$\lambda x.\texttt{blue}(x) \wedge \texttt{cup}(x) \wedge \texttt{comp.}(x) \wedge \texttt{screen}(x)$



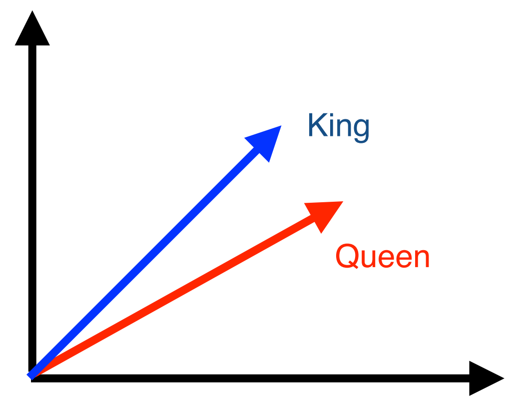
Two extremes on language understanding

Two extremes on the language understanding



$$z \sim p_\theta(z \mid \mathbf{x})$$

$$y = [\![z]\!]_w$$



T. Mikolov, et. al. "Linguistic Regularities in Continuous Space Word Representations" NAACL 2013