Jointly Learning to Parse and Perceive: Connecting Natural Language to the Physical World

Jayant Krishnamurthy

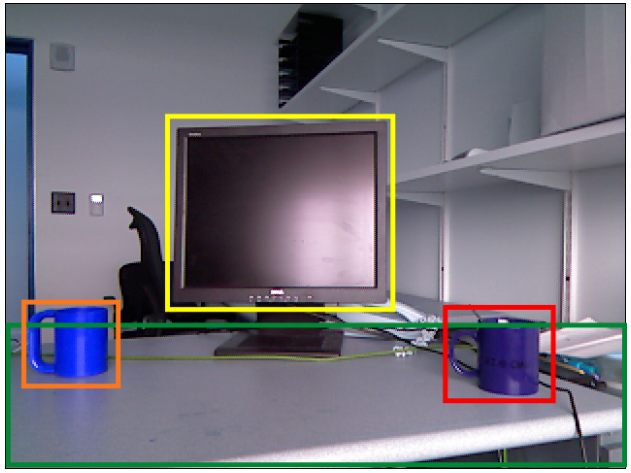
(Joint work with Thomas Kollar)

Grounded Language Understanding

"Go get my pen from my office"

Grounding Natural Language Descriptions

Input:

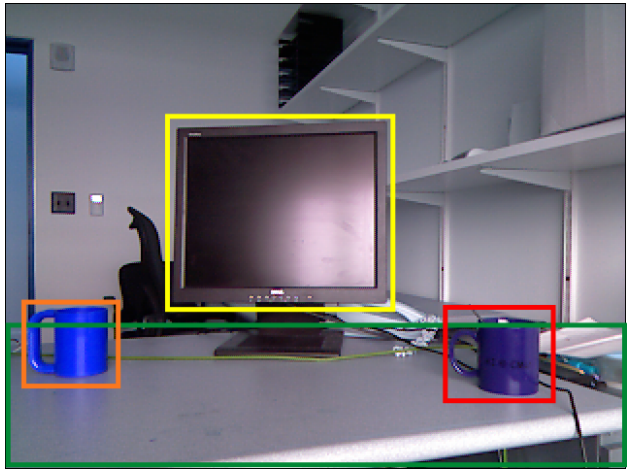


"The mugs."

Output:

Grounding Natural Language Descriptions

Input:



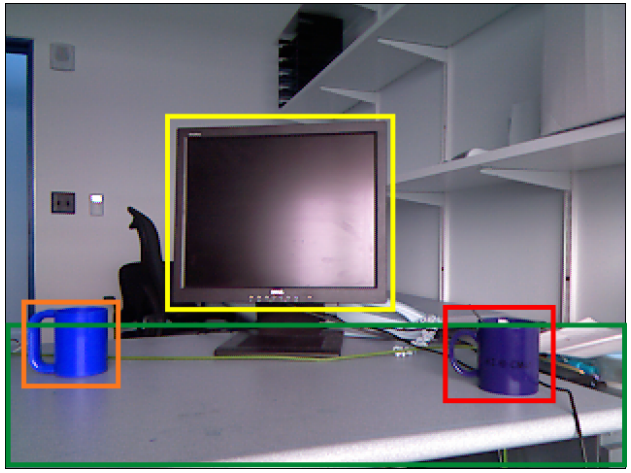
"The mugs."

"The mug to the left of the monitor."

Output:

Grounding Natural Language Descriptions

Input:



"The mugs."

"The mug to the left of the monitor."

Output:

- Set-valued output (unlike: [Kollar et al., 2010] [Tellex et al., 2011])
- No *a priori* knowledge base (unlike: [Zelle & Mooney 1996] [Zettlemoyer & Collins 2005] [Chen & Mooney, 2008,2011] [Liang et al., 2011])
- Relational language (unlike: [Matuszek et al., 2012])

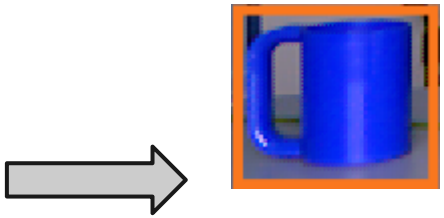
Outline

- Introduction
- Logical Semantics with Perception (LSP)
- Weakly-Supervised Training
- Experiments

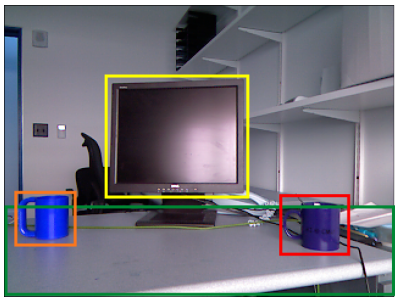
Logical Semantics with Perception (LSP)

"There's a mug to the left of the monitor"

Language



Output



Environment

Logical Semantics with Perception (LSP)

"There's a mug to the left of the monitor"

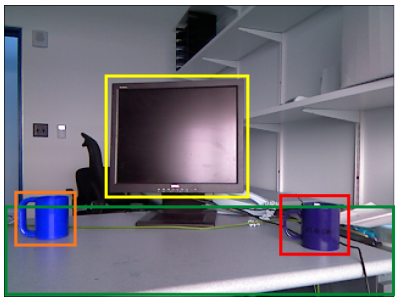
Language

Parsing

$\lambda x. \exists y. mug(x) \wedge left(x,y) \wedge monitor(y)$

Semantic Parse

Output



Environment

[Zettlemoyer & Collins, 2005]

Logical Semantics with Perception (LSP)

"There's a mug to the left of the monitor"

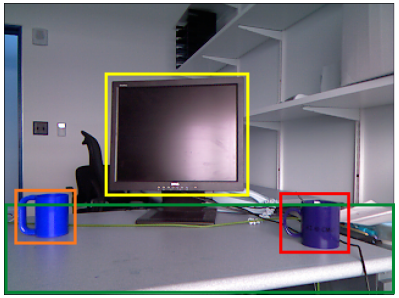
Language

Parsing

$\lambda x. \exists y. mug(x) \wedge left(x,y) \wedge monitor(y)$

Semantic Parse

Output



Environment

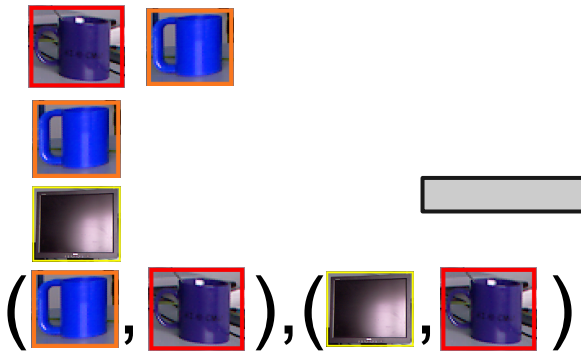
Perception

mug

blue

monitor

left



Knowledge Base

Logical Semantics with Perception (LSP)

"There's a mug to the left of the monitor"

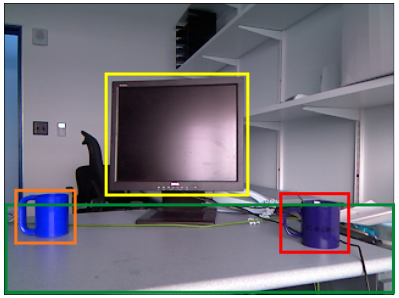
Language

Parsing

$\lambda x. \exists y. mug(x) \wedge left(x,y) \wedge monitor(y)$

Semantic Parse

Output



Environment

Perception

mug

blue

monitor

left

(,), (,)

Knowledge Base

Logical Semantics with Perception (LSP)

"There's a mug to the left of the monitor"

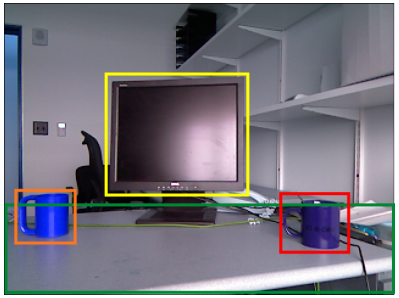
Language

Parsing

$\lambda x. \exists y. mug(x) \wedge left(x,y) \wedge monitor(y)$

Semantic Parse

Output



Environment

Perception

mug

blue

monitor

left

(,), (,)

Knowledge Base

LSP in Math

Parsing

$$\theta_{prs}^T \phi_{prs}(\ell, t, z)$$

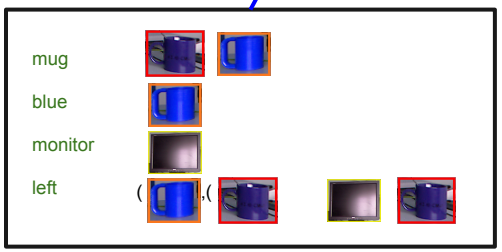
Perception

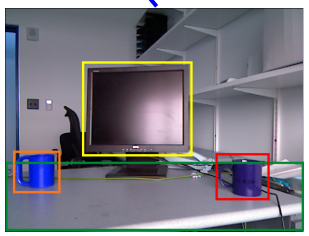
Evaluation (deterministic)

$$f(\gamma, \Gamma, \ell, t, z, d; \theta) = f_{prs}(\ell, t, z; \theta_{prs}) + f_{per}(\Gamma, d; \theta_{per}) + f_{eval}(\gamma, \Gamma, \ell)$$

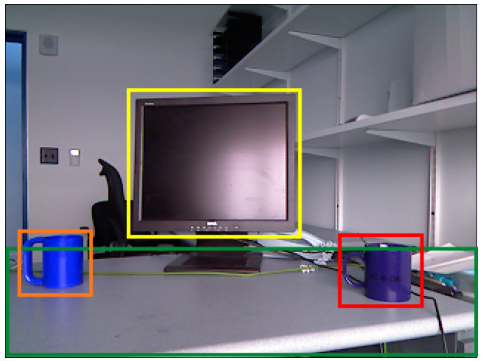
λx. ∃ y.mug(x) ∧ left(x,y) ∧ monitor(y)

"There's a mug to the left of the monitor"



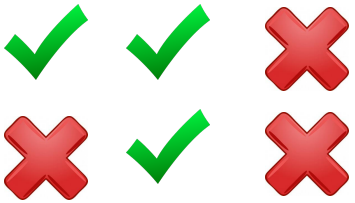


Perception



mug

blue



Per-predicate classifiers:

$\theta^T_{mug}\phi_{cat}($  $)$

$\theta^T_{left}\phi_{rel}($  ,  $)$

Outline

- Motivation
- Logical Semantics with Perception (LSP)
- Weakly-Supervised Training
- Experiments

Weakly-Supervised Training

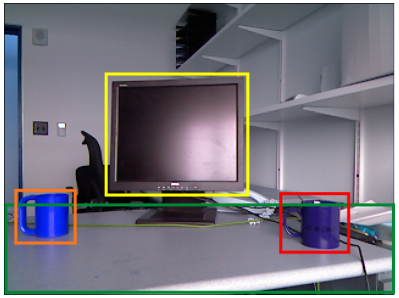
"There's a mug to the left of the monitor"

Language

?

Semantic Parse

Output

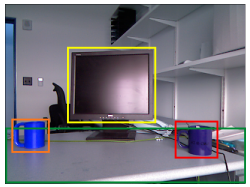


Real World

?

Knowledge Base

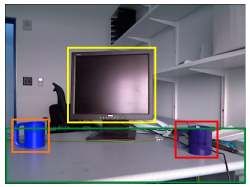
Parameter Estimation

$d^i =$ 

$z^i =$ "There's a mug to the left of the monitor"

$\gamma^i = \{$  $\}$

Parameter Estimation

$d^i =$ 

$z^i =$ "There's a mug to the left of the monitor"

$\gamma^i = \{$  $\}$

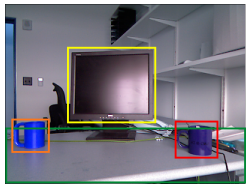
Best prediction for input

$$\hat{\gamma}, \hat{\Gamma}, \hat{\ell}, \hat{t} \leftarrow \arg\max_{\gamma,\Gamma,\ell,t} f(\gamma, \Gamma, \ell, t, z^i, d^i; \theta^j) + \text{cost}(\gamma, \gamma^i)$$

Best explanation for correct answer

$$\Gamma^*, \ell^*, t^* \leftarrow \arg\max_{\Gamma,\ell,t} f(\gamma^i, \Gamma, \ell, t, z^i, d^i; \theta^j)$$

Parameter Estimation

$d^i =$  $z^i =$ "There's a mug to the left of the monitor" $\gamma^i = \{$  $\}$

Best prediction for input

$$\hat{\gamma}, \hat{\Gamma}, \hat{\ell}, \hat{t} \leftarrow \arg\max_{\gamma,\Gamma,\ell,t} f(\gamma, \Gamma, \ell, t, z^i, d^i; \theta^j) + \text{cost}(\gamma, \gamma^i)$$

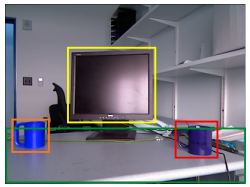
Best explanation for correct answer

$$\Gamma^*, \ell^*, t^* \leftarrow \arg\max_{\Gamma,\ell,t} f(\gamma^i, \Gamma, \ell, t, z^i, d^i; \theta^j)$$

Parser Update:

$$\theta^{j+1}_{prs} \leftarrow \theta^j_{prs} + \alpha_j \left(\phi_{prs}(\ell^*, t^*, z^i) - \phi_{prs}(\hat{\ell}, \hat{t}, z^i) \right)$$

Parameter Estimation

$d^i =$  $z^i =$ "There's a mug to the left of the monitor" $\gamma^i = \{$  $\}$

Best prediction for input

$$\hat{\gamma}, \hat{\Gamma}, \hat{\ell}, \hat{t} \leftarrow \arg\max_{\gamma,\Gamma,\ell,t} f(\gamma, \Gamma, \ell, t, z^i, d^i; \theta^j) + \text{cost}(\gamma, \gamma^i)$$

Best explanation for correct answer

$$\Gamma^*, \ell^*, t^* \leftarrow \arg\max_{\Gamma,\ell,t} f(\gamma^i, \Gamma, \ell, t, z^i, d^i; \theta^j)$$

Perception Update:

$\hat{\Gamma}$ mug ✓ ✗ ✓

Γ^*mug ✓ ✓ ✗

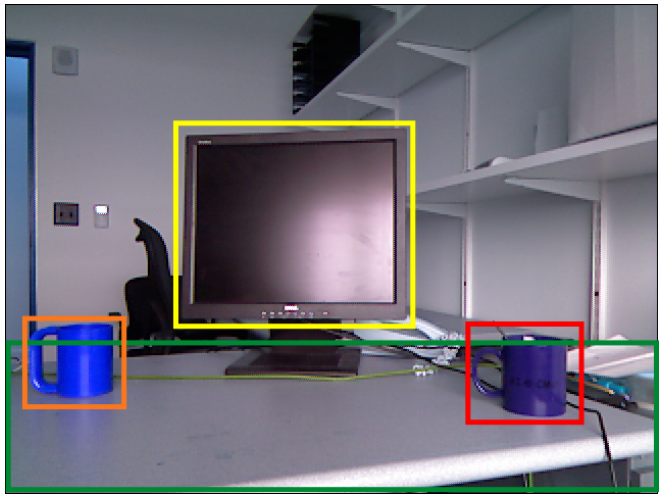
$$\theta^{j+1}_{mug} \leftarrow \theta^j_{mug} + \alpha_j \left(\phi_{cat}(\text{  }) - \phi_{cat}(\text{  })\right)$$

Outline

- Motivation
- Logical Semantics with Perception (LSP)
- Weakly-Supervised Training
- Experiments

Data sets

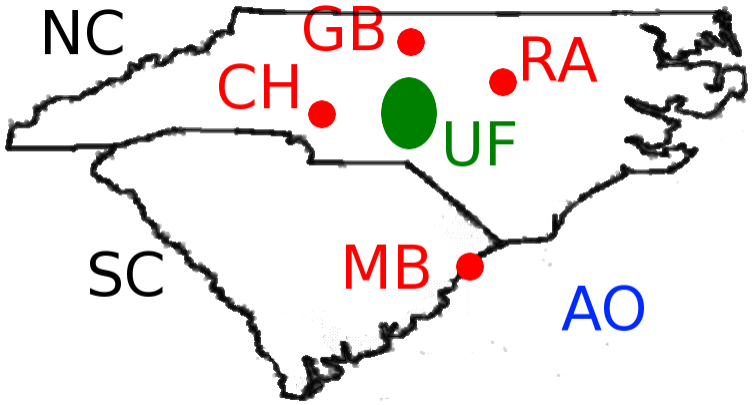
Scene



"A blue colored coffee mug is placed very near to the computer on the table."

(15 images, 284 natural language descriptions)

GeoQA



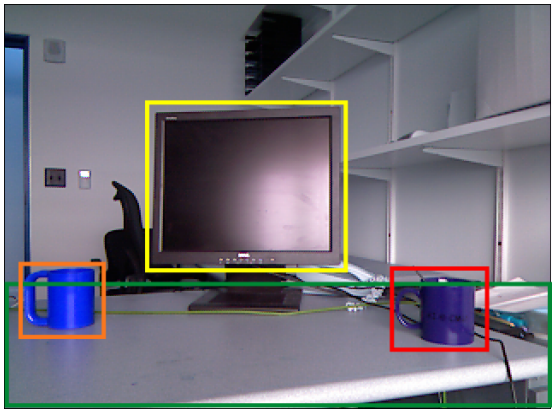
"What cities are east of Greensboro in North Carolina?"

RA (Raleigh)

(10 environments, 263 natural language questions)

Evaluation Metric

Test Example



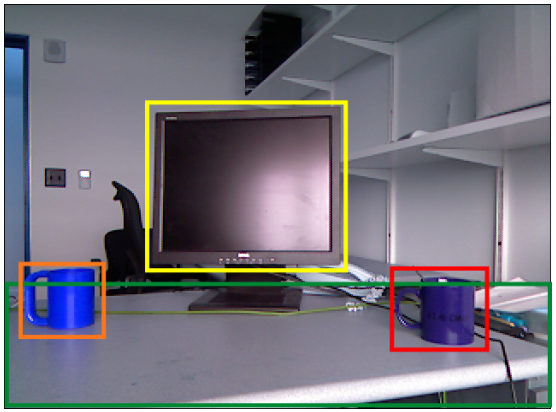
"There's a mug to the left of the monitor"

Prediction

Correct?

Evaluation Metric

Test Example



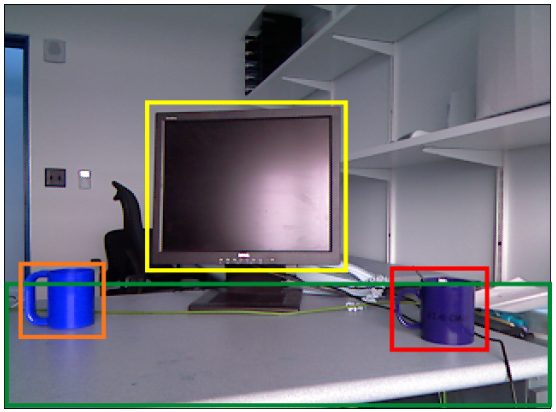
"There's a mug to the left of the monitor"

Prediction

Correct?

Evaluation Metric

Test Example



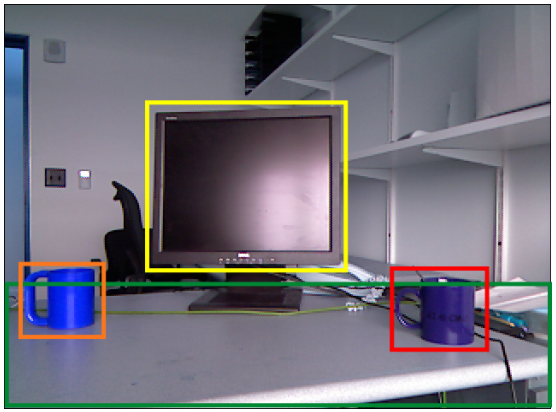
"There's a mug to the left of the monitor"

Prediction

Correct?

Evaluation Metric

Test Example

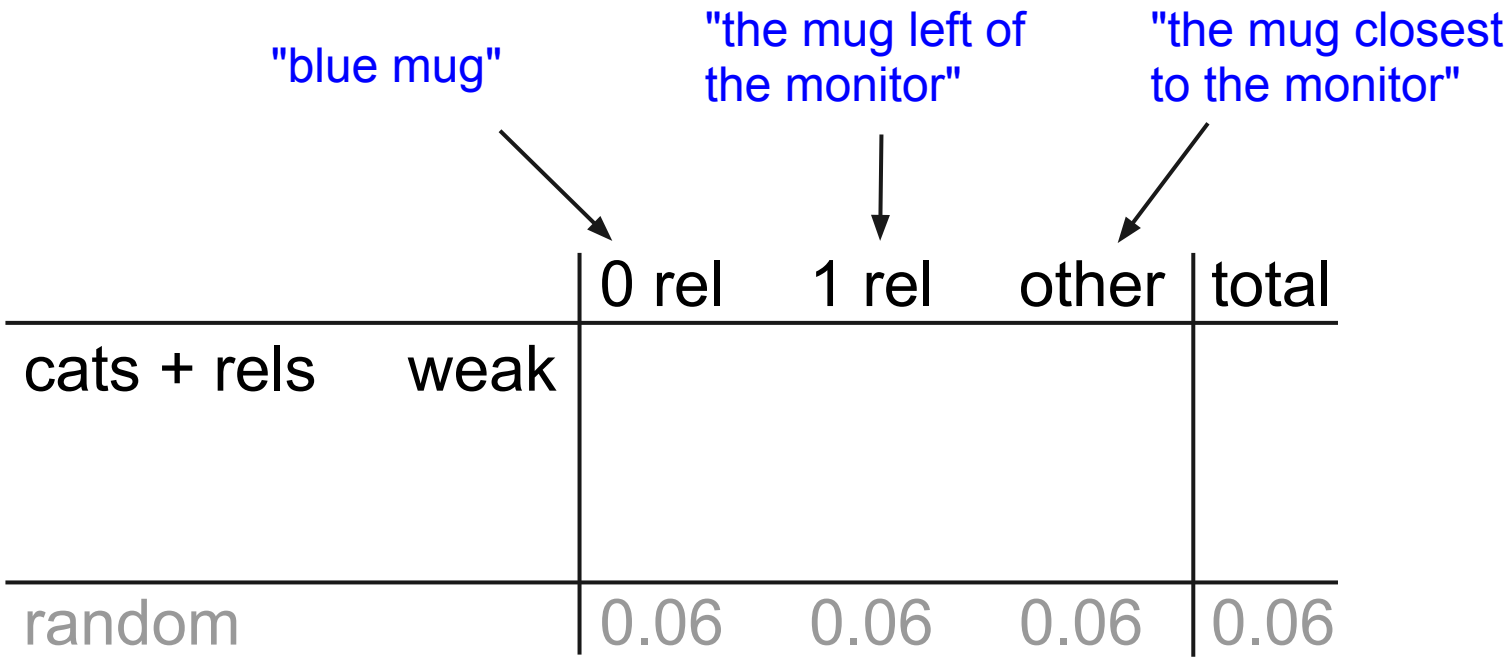


"There's a mug to the left of the monitor"

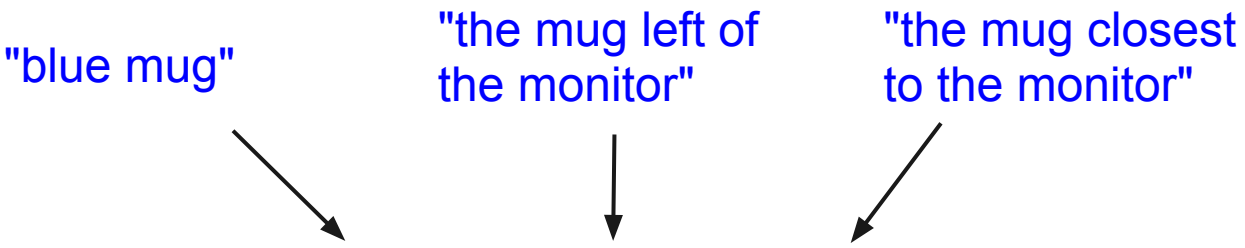
Prediction

Correct?

Results (Scene)



Results (Scene)



		0 rel	1 rel	other	total
cats + rels	weak	0.89	0.77	0.16	0.67
random		0.06	0.06	0.06	0.06

Results (Scene)

"blue mug"

"the mug left of the monitor"

"the mug closest to the monitor"

		0 rel	1 rel	other	total
cats + rels	weak	0.89	0.77	0.16	0.67
cats only	weak	0.94	0.45	0.20	0.51
random		0.06	0.06	0.06	0.06

[Matuszek et al., 2012]

Results (Scene)

"blue mug"

"the mug left of the monitor"

"the mug closest to the monitor"

		0 rel	1 rel	other	total
cats + rels	weak	0.89	0.77	0.16	0.67
cats only	weak	0.94	0.45	0.20	0.51
cats + rels	full	0.89	0.81	0.20	0.70
random		0.06	0.06	0.06	0.06

Results (GeoQA)

"what cities are there?"

"what states are south of north carolina?"

"what city is in eastern Tennessee?"

		0 rel	1 rel	other	total
cats + rels	weak	0.64	0.58	0.21	0.51
cats only	weak	0.22	0.12	0.07	0.17
cats + rels	full	0.64	0.53	0.21	0.48
random		0.01	0.01	0.01	0.01

Contributions

- Logical Semantics with Perception (LSP)
- Weakly-supervised training procedure
- Data available online: http://rtw.ml.cmu.edu/tacl2013_lsp/

Grounding Features

- Category Features

ϕ_{cat} () ⟶ Histogram of Oriented Gradients, Color Histogram

- Relation Features

ϕ_{rel} ( , ) ⟶ Spatial relation features