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

Language

Output

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

\[ \lambda x. \exists y. \text{mug}(x) \land \text{left}(x,y) \land \text{monitor}(y) \]

[Zettlemoyer & Collins, 2005]
"There's a mug to the left of the monitor"

\[ \lambda x. \exists y. \text{mug}(x) \land \text{left}(x, y) \land \text{monitor}(y) \]

**Language**

**Parsing**

**Semantic Parse**

**Output**

**Perception**

**Environment**

**Knowledge Base**
"There's a mug to the left of the monitor"
"There's a mug to the left of the monitor"

\[ \lambda x. \exists y. \text{mug}(x) \land \text{left}(x,y) \land \text{monitor}(y) \]
LSP in Math

**Parsing**

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

**Perception**

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

\[ \lambda x. \exists y. \text{mug}(x) \land \text{left}(x, y) \land \text{monitor}(y) \]

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

Per-predicate classifiers:

\[ \theta^T_{mug} \phi_{cat}( ) \quad \theta^T_{left} \phi_{rel}( ) \]
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

\[ d^i = \quad z^i = \text{"There's a mug to the left of the monitor"} \quad \gamma^i = \{ \text{mug} \} \]
Parameter Estimation

\[ \hat{d}^i = \overset{\text{Best prediction}}{\underset{\gamma, \Gamma, \ell, t}{\arg \max}} f(\gamma, \Gamma, \ell, t, z^i, d^i; \theta^j) + \text{cost}(\gamma, \gamma^i) \]

\[ \hat{z}^i = \text{"There's a mug to the left of the monitor"} \]

\[ \gamma^i = \{ \text{mug} \} \]

Best prediction for input

Best explanation for correct answer
Parameter Estimation

\[ d^i = \quad z^i = \quad \gamma^i = \{ \}

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

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

\[ d^i \quad z^i \quad \gamma^i \]

Best prediction for input

\[ \hat{\gamma}, \hat{\Gamma}, \hat{l}, \hat{t} \leftarrow \text{arg max}_{\gamma, \Gamma, l, t} f(\gamma, \Gamma, l, t, z^i, d^i; \theta^j) + \text{cost}(\gamma, \gamma^i) \]

Best explanation for correct answer

\[ \Gamma^*, l^*, t^* \leftarrow \text{arg max}_{\Gamma, l, t} f(\gamma^i, \Gamma, l, t, z^i, d^i; \theta^j) \]

Perception Update:

\[ \hat{\Gamma} \quad \text{mug} \quad \checkmark \quad \times \quad \checkmark \quad \Gamma^* \text{mug} \quad \checkmark \quad \checkmark \quad \times \]

\[ \theta_{mug}^{j+1} \leftarrow \theta_{mug}^j + \alpha_j \left( \phi_{\text{cat}}(\text{mug}) - \phi_{\text{cat}}(\text{gray}) \right) \]
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?"

(10 environments, 263 natural language questions)
"There's a mug to the left of the monitor"
Evaluation Metric

Test Example

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

Prediction

Correct?
"There's a mug to the left of the monitor"
Evaluation Metric

Test Example

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

Prediction

Correct?

- Yes
- No
Results (Scene)

<table>
<thead>
<tr>
<th></th>
<th>0 rel</th>
<th>1 rel</th>
<th>other</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>cats + rels weak</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>random</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

"blue mug"
"the mug left of the monitor"
"the mug closest to the monitor"
# Results (Scene)

<table>
<thead>
<tr>
<th></th>
<th>0 rel</th>
<th>1 rel</th>
<th>other</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>cats + rels</strong></td>
<td>weak</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;blue mug&quot;</td>
<td>0.89</td>
<td>0.77</td>
<td>0.16</td>
<td>0.67</td>
</tr>
<tr>
<td>&quot;the mug left of the monitor&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;the mug closest to the monitor&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>random</strong></td>
<td></td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>
## Results (Scene)

<table>
<thead>
<tr>
<th></th>
<th>0 rel</th>
<th>1 rel</th>
<th>other</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>cats + rels</td>
<td>weak</td>
<td>0.89</td>
<td>0.77</td>
<td>0.16</td>
</tr>
<tr>
<td>cats only</td>
<td>weak</td>
<td>0.94</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>random</td>
<td></td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

- "blue mug"
- "the mug left of the monitor"
- "the mug closest to the monitor"

[Matuszek et al., 2012]
## Results (Scene)

<table>
<thead>
<tr>
<th></th>
<th>0 rel</th>
<th>1 rel</th>
<th>other</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>cats + rels</td>
<td>weak</td>
<td>0.89</td>
<td>0.77</td>
<td>0.16</td>
</tr>
<tr>
<td>cats only</td>
<td>weak</td>
<td>0.94</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>cats + rels</td>
<td>full</td>
<td>0.89</td>
<td>0.81</td>
<td>0.20</td>
</tr>
<tr>
<td>random</td>
<td></td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
</tbody>
</table>

- "blue mug"  
- "the mug left of the monitor"  
- "the mug closest to the monitor"
## Results (GeoQA)

<table>
<thead>
<tr>
<th>Question</th>
<th>0 rel</th>
<th>1 rel</th>
<th>other</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>cats + rels, weak</td>
<td>0.64</td>
<td>0.58</td>
<td>0.21</td>
<td>0.51</td>
</tr>
<tr>
<td>cats only, weak</td>
<td>0.22</td>
<td>0.12</td>
<td>0.07</td>
<td>0.17</td>
</tr>
<tr>
<td>cats + rels, full</td>
<td>0.64</td>
<td>0.53</td>
<td>0.21</td>
<td>0.48</td>
</tr>
<tr>
<td>random</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Contributions

● Logical Semantics with Perception (LSP)

● Weakly-supervised training procedure

● Data available online: http://rtw.ml.cmu.edu/tacl2013_lsp/
Grounding Features

- **Category Features**

  \( \phi_{cat} (\text{ }) \)  →  Histogram of Oriented Gradients, Color Histogram

- **Relation Features**

  \( \phi_{rel} (\text{ }, \text{ }) \)  →  Spatial relation features