Natural Language Communication with Robots

Yonatan Bisk
ISI-USC

Joint work with:

Deniz Yuret
Koç University

Daniel Marcu
ISI-USC
Components of Communication

Entity/Spatial Grounding

Understanding

Planning and Plan Recognition

Language Generation

....
Grounding

The third block from the left
place the **nvidia** block **east** of the **hp** block.
Plans

Draw the number six with a rigid base and a right diagonal top. Start with a line of 6 blocks in the middle of the table ...
Generation

[I need to] move UPS from the left side of the board to just below Starbucks, leaving a small gap.
Goal

Introduce a dataset collection paradigm for

**Human-Robot Communication:**
Understanding, Learning, and Generation

1. Easily evaluated
2. Data exists in 3D space
3. Natural language utterances
4. Parallel annotation at differing levels of abstraction
5. Computer Vision can help but is not a pre-requisite

+ Models to begin addressing understanding
Dataset
Action Sequences

Identifiable Sequences

Random Blank Sequences
Problem Solution Sequences

0
1
13
14
20

Single
Short Seq

Single

Long Seq

We focus on Single Actions in this work
Corpus Creation

Simple Actions

*Move HP in front of Twitter and slightly to the left*
Remove the block above the right bottom block and place it on top of the left stack of blocks.
Nine Annotations

1. coca cola, hp, nvidia.
2. nvidia, to the right of hp
3. place the nvidia block east of the hp block.
4. move the nvidia block to the right of the hp block
5. place the nvidia block to the east of the hp block.
6. move the nvidia block directly to the right of the hp block.
7. move the nvidia block just to the right of the hp block in line with the mercedes block.
8. put the nvidia block on the right end of the row of blocks that includes the coca cola and hp blocks.
9. put the nvidia block on the same row as the coca cola block, in the first open space to the right of the coca cola block.
## Corpus Statistics V1

<table>
<thead>
<tr>
<th></th>
<th>Actions</th>
<th>Types</th>
<th>Tokens</th>
<th>Ave Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>11,870</td>
<td>1,359</td>
<td>~257K</td>
<td>15 tokens</td>
</tr>
<tr>
<td>Random</td>
<td>2,492</td>
<td>1,172</td>
<td>~84K</td>
<td>23.5 tokens</td>
</tr>
</tbody>
</table>
Natural Language Understanding
Action Understanding

Given:
World
Utterance

Goal:
Execute a command

Block to Move
\((x, y, z)_S\)

Where to Move
\((x, y, z)_T\)

place the **nvidia** block east of the **hp** block.
World Representation

Images (w/ Occlusion)

Exact Locations

<table>
<thead>
<tr>
<th>Brand</th>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adidas</td>
<td>0.8</td>
<td>0.1</td>
<td>0.76</td>
</tr>
<tr>
<td>BMW</td>
<td>-0.3</td>
<td>0.1</td>
<td>-0.4</td>
</tr>
<tr>
<td>Burger King</td>
<td>0.5</td>
<td>0.1</td>
<td>0.14</td>
</tr>
<tr>
<td>Coke</td>
<td>-0.07</td>
<td>0.1</td>
<td>0.00</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This Work
20 x 3 Matrix
Evaluation: Euclidean Distance

Block to Move

$$|| (x, y, z)_{SPred} - (x, y, z)_{SGold} ||_2$$

Where to Move

$$|| (x, y, z)_{TPred} - (x, y, z)_{TGold} ||_2$$
Baseline Models

Output:

Block to Move

\((x, y, z)_S\)

Where to Move

\((x, y, z)_T\)

Random

Random Block to move
Random Block to place it next to

Center

Perfect knowledge of which block to move
Always place it in the center of the board

We also Perform Human Evaluation
Simple Semantics

Model 1: A Discrete world (Source, Direction, Reference)

Move the **BMW** block *in front of* the **Adidas** block

Move the **Source** block **Direction** the **Reference** block

\[\begin{array}{c|c|c}
\text{NW} & \text{N} & \text{NE} \\
\hline
\text{W} & \text{TOP} & \text{E} \\
\hline
\text{SW} & \text{S} & \text{SE} \\
\end{array}\]

\[\begin{array}{c}
\in [1,20] \\
\in [1,9] \\
\in [1,20]
\end{array}\]
Simple Semantics

Model 1: A Discrete world (Source, Direction, Reference)

- Embedding
- FF
- Softmax

Forced Semantic Structure

- Source: \[1, 20\] Block IDs
- Direction: \[1, 9\] Block IDs
- Target: \[1, 20\] Block IDs

(S,D,R) programatic conversion to (x,y,z)
End-to-End Model

Move the BMW block in front of the Adidas block

\[(x, y, z)_{SPred}\]

or

\[(x, y, z)_{TPred}\]
End-to-End Model

Move the **BMW** block **in front of** the **Adidas** block

Assumed Logic: Can we encode this?

Direction

$\pm x, \pm y, \pm z$

Reference

$(x, y, z)$

$(x, y, z)_{TPred}$
End-to-End Model

Encoder

Representation

Grounding

Prediction

W_1

... Hidden

... Hidden

W_n

Hidden

Semantics 1

Semantics 2

Semantics 3

Hidden

World (3x20)

(x, y, z)

Trained Twice
Source + Target
## MNIST Performance

<table>
<thead>
<tr>
<th>Source Method</th>
<th>Source Mean</th>
<th>Target Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>0.00</td>
<td>0.53</td>
</tr>
<tr>
<td>Simple Semantics</td>
<td><strong>0.14</strong></td>
<td><strong>0.98</strong></td>
</tr>
<tr>
<td>End-To-End</td>
<td>0.19</td>
<td>1.05</td>
</tr>
<tr>
<td>Center Baseline</td>
<td></td>
<td>3.43</td>
</tr>
<tr>
<td>Random Baseline</td>
<td>6.49</td>
<td>6.21</td>
</tr>
</tbody>
</table>
## Blank Block Performance

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human</strong></td>
<td>0.30</td>
</tr>
<tr>
<td><strong>Simple Semantics</strong></td>
<td>5.00</td>
</tr>
<tr>
<td><strong>End-To-End</strong></td>
<td>3.47</td>
</tr>
<tr>
<td><strong>Center Baseline</strong></td>
<td>4.06</td>
</tr>
<tr>
<td><strong>Random Baseline</strong></td>
<td>4.97</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human</strong></td>
<td>1.39</td>
</tr>
<tr>
<td><strong>Simple Semantics</strong></td>
<td>5.57</td>
</tr>
<tr>
<td><strong>End-To-End</strong></td>
<td>3.70</td>
</tr>
<tr>
<td><strong>Center Baseline</strong></td>
<td>4.06</td>
</tr>
<tr>
<td><strong>Random Baseline</strong></td>
<td>5.44</td>
</tr>
</tbody>
</table>
Common Errors

**Multi-relation actions**
Place block 20 parallel with the 8 block and slightly to the right of the 6 block.

**Geometric Understanding**
Continue the diagonal row of 20, 19 and 15 downward with 13.

**Grammatical Ambiguity**
19 moved from behind the 8 to under the 18th block.
Summary

This Work:

• Initial Models for Language Understanding

• An environment for exploring grounded phenomena

Moving Forward:

• Language Generation, Planning, …

• Increased task difficulty.
Thanks!

http://nlg.isi.edu/language-grounding/