Linguistic scaffolds for policy learning

Jacob Andreas
Berkeley → Microsoft Semantic Machines → MIT
Linguistic scaffolds for policy learning

(what can language do for RL?)

Jacob Andreas
Berkeley → Microsoft Semantic Machines → MIT
An NLPPer’s view of RL

$$\left( X, R \right) \rightarrow$$
An NLPPer’s view of RL

$(\text{agent, } R) \rightarrow$ memorize 1 reward fn
An NLPer’s view of RL

\[(T, R) \rightarrow \text{(memorize k reward fns)}\]

[1020x349] memorize k reward fns

[e.g. Taylor & Stone 09]
An NLPer’s view of RL

Learn to accomplish new goals!

[e.g. Schaul et al. 15]
An NLPer’s view of RL

Learn to follow instructions!

- run northwest
- go southwest
Instructions as observations

((, R) → )

((, R₁) → )
((, R₂) → )

((, R₁) → )
((-2, 3) → )
((-2, -2) → )

((, R₁) → )
((, R₁) → )
((, R₁) → )

run northwest

go southwest
Instructions as observations

\((x, R)\) →

\((-2, 3)\)

\((-2, -2)\)

\((x, R_1)\) →

run northwest

\((x, R_1)\) →

go southwest
(1) Instructions are moves in a game, not observations of an environment.
Beyond goals

(2) There’s more to language learning than instruction following!

(???, \(R_1\)) → (???, \(R_1\))

(not so fast, \(R_1\)) → (???, \(R_1\))

(run northwest, \(R_1\)) → (???, \(R_1\))

(go southwest, \(R_1\)) → (???, \(R_1\))
Language use as gameplay
Turn right and walk through the kitchen. Go right into the living room and stop by the rug.

[Anderson et al. 18]
A reference game

[Frank & Goodman 12]
“glasses"
“glasses”
“glasses”
“glasses"
The rational speech acts model

\[ L_0(. \mid \text{glasses}) \]
\[ L_0(. \mid \text{hat}) \]

[Frank & Goodman 12, Degen 13]
The rational speech acts model

\[ L_0(. \mid \text{glasses}) \]
\[ L_0(. \mid \text{hat}) \]
\[ S_1(\text{glasses} \mid .) \propto L_0(. \mid \text{glasses}) \]
\[ S_1(\text{hat} \mid .) \]

[Frank & Goodman 12, Degen 13]
The rational speech acts model

\[ L_1(. \mid \text{glasses}) \propto S_1(\text{glasses} \mid .) \]
\[ L_1(. \mid \text{hat}) \]

\[ S_1(\text{glasses} \mid .) \propto L_0(. \mid \text{glasses}) \]
\[ S_1(\text{hat} \mid .) \]

\[ 3/4 \quad 0 \quad 1 \]
\[ 1/4 \quad 1 \quad 1/3 \]

[Frank & Goodman 12, Degen 13]
Q: Do you know what time it is?
Pragmatics

Q: Do you know what time it is?
A: Yes
Pragmatics

Q: Do you know what time it is?
   A: Yes

I find his cooking very interesting.

[Grice 70]
RSA game tree

speaker

hat

glasses
RSA game tree: as speaker

speaker

hat

hat

+1

-1

glasses

glasses

+1

-1

listener
RSA game tree: as speaker

speaker

hat

hat

glasses

glasses

listener

+1

-1

+1

-1
RSA game tree: as listener

- Speaker
- Listener

- Glasses

?-?

?-?

?-?
A recipe for pragmatic language understanding

1. Train a base **speaker** model
A recipe for pragmatic language understanding

1. Train a base **speaker** model

2. Solve this POMDP:

Application: instruction following

human: Go through the door on the right and continue straight. Stop in the next room in front of the bed.
Application: instruction generation

**seq2seq:** Walk past the dining room table and chairs and wait there.

**reasoning:** Walk past the dining room table and chairs and take a right into the living room. Stop once you are on the rug.

**human:** Turn right and walk through the kitchen. Go right into the living room and stop by the rug.
Utterances are chosen to facilitate correct interpretation in context.

(This makes the learning problem easier!)
Language as a scaffold for learning
What else is an instruction follower good for?

- Language learning
  - go east of the heart

- Reinforcement learning

Pretraining via language learning

π(⟨; η⟩) → NORTH

go east of the heart

[Branavan et al., 09]
(Standard) reinforcement learning
Concept learning

$R(\pi(\cdot; \eta, \circ), \rightarrow) \rightarrow \text{NORTH, ...}$

find the horse
Concept learning

\[ R(\pi(\cdot; \eta, \_)) \rightarrow \text{NORTH, ...} \]

-0.52

find the horse
Concept learning

\[ R(\pi(\text{left of heart}, \eta), \cdot) \rightarrow \text{SOUTH, ...} \]

-0.52

find the horse

0.33

left of heart
Concept learning

\[ R(\pi(; \eta, \mathcal{C})), \rightarrow \text{SOUTH, ...} \]

- find the horse: -0.52
- left of the heart: 0.33
- heart east side: 0.95
As multitask learning

Language learning

\[
\arg \min_{\eta} R(\pi(\hat{\nabla} \mid \text{blue}; \eta, \cdot))
\]

Reinforcement learning

\[
\arg \min_{\eta} R(\pi(\hat{\nabla} \mid \text{green}; \eta, \cdot))
\]

[Caruana 97]
As a language game...

$$\text{arg min}$$

$$\quad \quad \quad \quad \quad \quad \pi \circ R$$

speaker model  

$\text{go east of the heart}$

listener loss

$-0.52$
Examples: Navigation

Human description:
- move to the star

Inferred description:
- reach the star cell
- reach square one right of triangle
- reach cell to the right of the triangle
- reach cell on left of triangle
- reach square left of triangle
- reach spade
- go to the spade
- left of the circle

True description

Pred description

reach cell on left of triangle

reach square left of triangle
Results: RL

This work
Multitask
Scratch

Average reward
Timestep (×1000)
20 40 60 80 100

Figure 8: Learning curves for treasure hunting. These results show the average reward obtained by each learning algorithm across multiple evaluation environments, after language learning has already taken place. The learner is trained from task-specific evidence intervals for mean performance. The max possible reward for this task is 3 points. Error bands shows 95% confidence comes from fine-tuning. In the case of RL in particular, the contribution from \( L_3 \) is effectively an in-place adaptation following model of a kind well-studied in orthogonal to those of meta-learning—one could imagine using annotations are in this case adapted from a natural language learning has already taken place.

To learn, we sample a fixed number of descriptions more efficiently, or use a technique like RL requiring them to immediately perform well on held-out environments, rather than zero-shot reward as the training criterion for the inverting the language-learning phase, and adapts to individual environments using "vanilla" policy gradient algorithms. In reinforcement learning, we typically encourage concept-learning observations. Here, agents are free to interact with the environment as much as they need, but receive observations only during interaction. Why should we expect \( L_3 \) to help in this setting?
## Results

### Examples

<table>
<thead>
<tr>
<th>emboldens</th>
<th>emboldecs</th>
</tr>
</thead>
<tbody>
<tr>
<td>kisses</td>
<td>kisses</td>
</tr>
<tr>
<td>loneliness</td>
<td>locelices</td>
</tr>
<tr>
<td>vein</td>
<td>veic</td>
</tr>
<tr>
<td>dogtrot</td>
<td>dogtrot</td>
</tr>
</tbody>
</table>

### True Description

- replace all n s with c

### True Output

- loocies

### Pred. Description

- change any n to a c

### Pred. Output

- loocies

---

6 **Policy Search**

The previous two sections examined supervised settings where the learning signal comes from few examples but is readily accessible. In this section, we move to a set of reinforcement learning problems, where the learning signal is instead sparse and time-consuming to obtain. We evaluate on a collection of 2-D treasure hunting tasks. These tasks require the agent to discover a rule that determines the location of buried treasure in a large collection of environments of the kind shown in Figure 7. To recover the treasure, the agent must navigate (while avoiding water) to its goal location, then perform a DIG action. At this point the episode ends; if the treasure is located in the agent's current position, it receives a reward, otherwise it does not. In every task, the treasure has consistently been buried at a fixed position relative to some landmark (like the heart in Figure 7). Both the offset and the identity of the target landmark are unknown to the agent, and the location landmark itself varies across maps. Indeed, there is nothing about the agent's observations or action space to suggest that landmarks and offsets are even the relevant axis of variation across tasks, but this structure is made clear in the natural language annotations. The high-level structure of these tasks is similar to one used by Hermer-Vazquez et al. (2001) to study concept learning in humans.

The interaction between language and learning in these tasks is rather different than in the supervised settings. In the supervised case, language served mostly as a guard against overfitting, and could...
Results: programming by demonstration

- Identity: 18
- Multitask: 50
- Meta: 62
- This Work: 76
Results: locomotion

Modular multitask reinforcement learning with policy sketches. A, Klein & Levine. ICML 2017

north, east, north
Generalization

This work

Multitask
Learning with corrections

Language learning

Reinforcement learning

go north a bit more
Pretraining by learning to correct

\[ \pi(\cdot; \eta, \cdot) \rightarrow \text{NORTH} \]

Further east

Guiding policies with language via meta-learning. ICLR 19.
Pretraining by learning to correct

\[ \pi \left( \cdot \mid \eta, \right) \rightarrow \text{NORTH} \]

further east
Learning from corrections

\[ \pi(\cdot; \eta, \text{rectangle}) \rightarrow \text{WEST, ...} \]

\[ \pi(\cdot; \eta, \text{go to top}) \rightarrow \text{NORTH, ...} \]

\[ \pi(\cdot; \eta, \text{further west}) \rightarrow \text{NORTH, ...} \]
Touch cyan block.

Move closer to magenta block.

Move a lot up.

Move a little up.
Enter the blue room.
Enter the red room.
Exit the blue room.

Pick up the blue triangle.
Language is useful as side information, not just a goal specification.

Use it with / instead of instructions as a representational bottleneck or interactive advice.
So what comes next?
What comes next?

Challenges for the field:
What comes next?

Challenges for the field:

- huge datasets
What comes next?

Challenges for the field:

- **huge** datasets
- with **fake** annotations
What comes next?

Challenges for the field:

- **huge** datasets
- with **fake** annotations
- that look **very little like** natural language
What comes next?

Challenges for the field:

- **huge** datasets → Learn to make do without an annotation for every rollout!
- with **fake** annotations
- that look **very little like natural language**
What comes next?

Challenges for the field:

- **huge** datasets → Learn to make do without an annotation for every rollout!
- **fake** annotations → Learn to generalize from fake strings to real ones!
- that look **very little like natural language**
What comes next?

Challenges for the field:

- **huge** datasets  →  Learn to make do without an annotation for every rollout!
- **fake** annotations  →  Learn to generalize from fake strings to real ones!
- that look **very little like natural language**  →  Pay attention to human evals (or scope claims accordingly)!
Learn more: Luketina et al.,
A survey of reinforcement learning informed by natural language

Task-independent

- Pre-training
- $v_{key}$, $v_{skull}$, $v_{ladder}$, $v_{rope}$

<table>
<thead>
<tr>
<th>Role</th>
<th>Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task-independent</td>
<td></td>
<td>having the correct key</td>
</tr>
<tr>
<td></td>
<td></td>
<td>known lock and skull</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unless the correct key is inserted</td>
</tr>
</tbody>
</table>

Task-dependent

**Language-assisted**

**Key**

- Opens a door of the same color as the key.

**Skull**

- They come in two varieties, rolling skulls and bouncing skulls ... you must jump over rolling skulls and walk under bouncing skulls.

**Language-conditional**

- Go down the ladder and walk right immediately to avoid falling off the conveyor belt, jump to the yellow rope and again to the platform on the right.