Grounded language understanding

Christopher Potts

Stanford Linguistics

CS 224U: Natural language understanding
May 6
Overview

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Listeners: From language to the world
4. Grounded chat bots
5. Reasoning about other minds
6. A few other grounding ideas
In the 1967 Stanley Kubrick movie *2001: A Space Odyssey*, the spaceship’s computer HAL can

- display graphics;
- play chess; and
- conduct natural, open-domain conversations with humans.

How well did the filmmakers do at predicting what computers would be capable in 2001?

(Slide idea from Andrew McCallum)
HAL

Jurassic Park (1993)

(Slide idea from Andrew McCallum)
HAL

Chess

HAL

Deep Blue (1997)

(Slide idea from Andrew McCallum)
David Bowman: Open the pod bay doors, HAL.

HAL: I’m sorry, Dave, I’m afraid I can’t do that.

David: What are you talking about, HAL?

HAL: I know that you and Frank were planning to disconnect me, and I’m afraid that’s something I cannot allow to happen.
Siri

You: Any good burger joints around here?

Siri: I found a number of burger restaurants near you.

You: Hmm. How about tacos?

Apple: [Siri remembers that you asked about restaurants. so it will look for Mexican restaurants in the neighborhood. And Siri is proactive, so it will question you until it finds what you’re looking for.]

(Slide from Marie de Marneffe)
Siri

Colbert: For the love of God, the cameras are on, give me something?

Siri: What kind of place are you looking for? Camera stores or churches?

[...]

Colbert: I don’t want to search for anything! I want to write the show!

Siri: Searching the Web for “search for anything. I want to write the shuffle.”

(Slide from Marie de Marneffe)
Language is action

Winograd (1986:170):
“all language use can be thought of as a way of activating procedures within the hearer. We can think of an utterance as a program – one that indirectly causes a set of operations to be carried out within the hearer’s cognitive system.”
Levinson’s (2000) analogy

Figure 0.1
Rembrandt sketch
Levinson’s (2000) analogy

“She interpret this sketch instantly and effortlessly as a gathering of people before a structure, probably a gateway; the people are listening to a single declaiming figure in the center. […] But all this is a miracle, for there is little detailed information in the lines or shading (such as there is). Every line is a mere suggestion […]. So here is the miracle: from a merest, sketchiest squiggle of lines, you and I converge to find adumbration of a coherent scene […]."
Levinson’s (2000) analogy

“We interpret this sketch instantly and effortlessly as a gathering of people before a structure, probably a gateway; the people are listening to a single declaiming figure in the center. [...] But all this is a miracle, for there is little detailed information in the lines or shading (such as there is). Every line is a mere suggestion [...]. So here is the miracle: from a merest, sketchiest squiggle of lines, you and I converge to find adumbration of a coherent scene [...].

“The problem of utterance interpretation is not dissimilar to this visual miracle. An utterance is not, as it were, a veridical model or “snapshot” of the scene it describes [...]. Rather, an utterance is just as sketchy as the Rembrandt drawing.”
<table>
<thead>
<tr>
<th>Linguistic insights</th>
<th>Speakers</th>
<th>Listeners</th>
<th>Grounded chat bots</th>
<th>Other minds</th>
<th>Other ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indexicality</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. I am speaking.
2. We won. [A team I'm on; a team I support; ...]
3. I am here [classroom; Stanford; ... planet earth; ...]
4. We are here. [pointing at a map]
5. I'm not here now. [old-fashioned answering machine]
6. We went to a local bar after work.
7. three days ago, tomorrow, now
Indexicality

1. I am speaking.
## Indexicality

1. I am speaking.
2. We won. [A team I’m on; a team I support; . . . ]
Indexicality

1. I am speaking.
2. We won. [A team I’m on; a team I support; ...]
3. I am here [classroom; Stanford; ... planet earth; ...]
Indexicality

1. I am speaking.
2. We won. [A team I’m on; a team I support; ...]
3. I am here [classroom; Stanford; ... planet earth; ...]
4. We are here. [pointing at a map]
Indexicality

1. I am speaking.
2. We won. [A team I’m on; a team I support; ... ]
3. I am here [classroom; Stanford; ... planet earth; ... ]
4. We are here. [pointing at a map]
5. I’m not here now. [old-fashioned answering machine]
Indexicality

1. I am speaking.
2. We won. [A team I’m on; a team I support; . . . ]
3. I am here [classroom; Stanford; . . . planet earth; . . . ]
4. We are here. [pointing at a map]
5. I’m not here now. [old-fashioned answering machine]
6. We went to a local bar after work.
Indexicality

1. I am speaking.
2. We won. [A team I’m on; a team I support; ...]
3. I am here [classroom; Stanford; ... planet earth; ...]
4. We are here. [pointing at a map]
5. I’m not here now. [old-fashioned answering machine]
6. We went to a local bar after work.
7. three days ago, tomorrow, now
Context dependence

Where are you from?
Context dependence

Where are you from?

- Connecticut. (Issue: birthplaces)
Context dependence

*Where are you from?*

- *Connecticut.* (Issue: birthplaces)
- *The U.S.* (Issue: nationalities)
Context dependence

Where are you from?

- Connecticut. (Issue: birthplaces)
- The U.S. (Issue: nationalities)
- Stanford. (Issue: affiliations)
Context dependence

Where are you from?

- **Connecticut.** (Issue: birthplaces)
- **The U.S.** (Issue: nationalities)
- **Stanford.** (Issue: affiliations)
- **Planet earth.** (Issue: intergalactic meetings)
Context dependence

I didn’t see any.
Context dependence

- Are there typos in my slides?

*I didn’t see any.*
Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?

*I didn’t see any.*
Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?
- Are there cookies in the cupboard?

I didn’t see any.
Context dependence

- Are there typos in my slides?
- Are there bookstores downtown?
- Are there cookies in the cupboard?
- ... 

I didn’t see any.
Context dependence

1. The light is on. Chris must be in his office.
2. The Dean passed a new rule. Chris must be in his office.
Context dependence

*If kangaroos had no tails, they would fall over.*

Seems true
Context dependence

*If kangaroos had no tails, they would fall over.*

Seems true, but suppose they had jetpacks.
Context dependence

"These two books contain the sum total of all human knowledge" (@James_Kpatrick)
Context dependence

“These two books contain the sum total of all human knowledge” (@James_Kpatrick)
Perspectival expressions

Please, when using the stairs
Stay to the right when going up,
stay to the left when going down.
This will keep people from
running into each other.
Routine pragmatic enrichment

how big is the contextually restricted domain of students?
what’s the additional contextual restriction?

false for most students?

who’s the speaker?

Many students met with me yesterday.

what’s the time of utterance?

but perhaps many met with the speaker at other times?
**SHRDLU (Winograd 1972)**

Full-grounded system that parses the user’s input, maps it to logical form, interprets that logical form in its world, tries to take appropriate action, and generates responses.

http://hci.stanford.edu/winograd/shrdlu/

One project did succeed. Terry Winograd’s program SHRDLU could use English intelligently, but there was a catch: the only subject you could discuss was a micro-world of simulated blocks.
SHRDLU (Winograd 1972)

1. Person: WHAT DOES THE BOX CONTAIN?
2. Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
3. Person: WHAT IS THE PYRAMID SUPPORTED BY?
4. Computer: THE BOX.
5. Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
6. Computer: FOUR OF THEM.
7. Person: IS AT LEAST ONE OF THEM NARROWER THAN THE ONE WHICH I TOLD YOU TO PICK UP?
9. Person: IS IT SUPPORTED?
10. Computer: YES, BY THE TABLE.
Winograd sentences

(Winograd 1972; Levesque 2013; Wang et al. 2018)
Winograd sentences

1. The trophy doesn’t fit into the brown suitcase because it’s too **small**. What is too small? **The suitcase** / The trophy

(Winograd 1972; Levesque 2013; Wang et al. 2018)
Winograd sentences

1. The trophy doesn’t fit into the brown suitcase because it’s too **small**. What is too small? **The suitcase** / The trophy

2. The trophy doesn’t fit into the brown suitcase because it’s too **large**. What is too large? **The suitcase** / **The trophy**

(Winograd 1972; Levesque 2013; Wang et al. 2018)
Winograd sentences

1. The trophy doesn’t fit into the brown suitcase because it’s too **small**. What is too small?
   **The suitcase** / The trophy

2. The trophy doesn’t fit into the brown suitcase because it’s too **large**. What is too large?
   The suitcase / **The trophy**

3. The council refused the demonstrators a permit because they **feared** violence. Who **feared** violence?
   **The council** / The demonstrators

(Winograd 1972; Levesque 2013; Wang et al. 2018)
Winograd sentences

1. The trophy doesn’t fit into the brown suitcase because it’s too small. What is too small?
   The suitcase / The trophy

2. The trophy doesn’t fit into the brown suitcase because it’s too large. What is too large?
   The suitcase / The trophy

3. The council refused the demonstrators a permit because they feared violence. Who feared violence?
   The council / The demonstrators

4. The council refused the demonstrators a permit because they advocated violence. Who advocated violence?
   The council / The demonstrators

(Winograd 1972; Levesque 2013; Wang et al. 2018)
Situated word learning

Children learn word meanings

1. with incredible speed
2. despite relatively few inputs
3. by using cues from
   ▶ contrast inherent in the forms they hear
   ▶ social cues
   ▶ assumptions about the speaker’s goals
   ▶ regularities in the physical environment.

Frank et al. (2012); Frank & Goodman (2014)
Consequences for NLU

- Human children are the best agents in the universe at learning language, and they depend heavily on grounding.

- Problems that are intractable without grounding are solvable with the right kinds of grounding.

- Deep learning is a flexible toolkit for reasoning about different kinds of information in a single model, so it’s led to conceptual and empirical improvements in this area.

- We should seek out (and develop) data sets that include the right kind of grounding.
Speakers: From the world to language

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Listeners: From language to the world
4. Grounded chat bots
5. Reasoning about other minds
6. A few other grounding ideas
## Color describer: Task formulation and data

<table>
<thead>
<tr>
<th>Color</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>green</td>
<td></td>
</tr>
<tr>
<td>purple</td>
<td></td>
</tr>
<tr>
<td>grape</td>
<td></td>
</tr>
<tr>
<td>turquoise</td>
<td></td>
</tr>
<tr>
<td>moss green</td>
<td></td>
</tr>
<tr>
<td>pinkish purple</td>
<td></td>
</tr>
<tr>
<td>light blue grey</td>
<td></td>
</tr>
<tr>
<td>robin’s egg blue</td>
<td></td>
</tr>
<tr>
<td>british racing green</td>
<td></td>
</tr>
<tr>
<td>baby puke green</td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Example from the xkcd color dataset as released by McMahan & Stone (2015).
Color descriptor: Training with teacher forcing
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*

![Diagram](image_url)

- Encoder:
  - Color rep
  - Color embedding
  - 208.3, 60, 88.2

- Decoder:
  - $x_1$
  - $<s>$
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*

Diagram:

- **Encoder**
  - color embedding
  - color rep
  - 208.3, 60, 88.2

- **Decoder**
  - <s>
  - x₁
  - h₁
  - dark

**Details:**

- Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*
**Color describer: Training with *teacher forcing***

---

**Encoder**

- **color rep**
- **color embedding**
- **208.3, 60, 88.2**

---

**Decoder**

- **<s>**
- **x₁**
- **h₁**
- **dark**
- **light**

**Embeddings**

- **error signal**
- **predicted probability distribution over the vocab**
- **one-hot encoding for next word**
- **derived from x₁ and color rep, the initial hidden state**

---

**Diagram Notes**

- **Decoder** is connected to **Encoder** through **color rep**.
- **208.3, 60, 88.2** are color embeddings.
- **<s>** is a special symbol for start of sentence.
- **h₁** is the initial hidden state derived from color rep and x₁.
- **light** and **dark** have probability distributions.
- **error signal** is used for training with teacher forcing.
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*
Color describer: Training with teacher forcing
Color describer: Training with *teacher forcing*

**Encoder**
- Color embedding: 208.3, 60, 88.2
- Color rep

**Decoder**
- <s>
- x₁
- x₃₇
- light
- h₁
- h₂
- dark
- blue

[Diagram showing the flow of information from the encoder to the decoder with teacher forcing.]
Color describer: Training with *teacher forcing*

Encoder
- color rep
- color embedding
- 208.3, 60, 88.2

Decoder
- blue
- dark
- $h_1$
- $x_1$
- $<s>$
- light
- $x_{37}$
- $h_2$
- error signal
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*

**Encoder**
- color rep
- color embedding
- 208.3, 60, 88.2

**Decoder**
- dark
- h₁
- h₂
- h₃
- blue
- x₁
- x₃₇
- x₁₁
- light
- <s>
- blue
Color describer: Training with *teacher forcing*.

```
<table>
<thead>
<tr>
<th>Encoder</th>
<th>Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>color rep</td>
<td>dark</td>
</tr>
<tr>
<td>color embedding</td>
<td>blue</td>
</tr>
<tr>
<td>208.3, 60, 88.2</td>
<td>green</td>
</tr>
<tr>
<td>h_1</td>
<td>h_2</td>
</tr>
<tr>
<td>x_1</td>
<td>x_37</td>
</tr>
<tr>
<td>&lt;s&gt;</td>
<td>light</td>
</tr>
<tr>
<td>h_3</td>
<td>blue</td>
</tr>
</tbody>
</table>
```

- **Encoder**
  - color rep
  - color embedding
  - 208.3, 60, 88.2

- **Decoder**
  - dark
  - blue
  - green
  - h_1
  - h_2
  - h_3
  - x_1
  - x_37
  - x_11
  - <s>
  - light
  - blue
Color describer: Training with *teacher forcing*
Color describer: Prediction

Encoder

Decoder

color rep

color embedding

208.3, 60, 88.2
Color describer: Prediction

Encoder
- color embedding
- color rep
- 208.3, 60, 88.2

Decoder
<s>
Color describer: Prediction

Encoder

- color rep
- color embedding
- 208.3, 60, 88.2

Decoder

- x₁
- <s>
Color describer: Prediction

```
208.3, 60, 88.2
```

```
<s>
```

```
x_1
```

```
h_1
```

```
color rep
```

```
color embedding
```

Encoder

Decoder
Color describer: Prediction

Encoder

- color rep
- color embedding
- 208.3, 60, 88.2

Decoder

- dark
- $h_1$
- $x_1$
- $<s>$
Color describer: Prediction

Encoder

- color rep
- color embedding
- 208.3, 60, 88.2

Decoder

- dark
- h₁
- x₁
- <s>
- h₂
- x₂₀
- dark
Color describer: Prediction
Color describer: Prediction

Encoder
- color rep
- color embedding
- 208.3, 60, 88.2

Decoder
- dark
- blue
- h₁
- h₂
- h₃
- x₁
- x₂₀
- x₁₁
- <s>
- dark
- blue
Color desriber: Prediction

Encoder

Decoder

color embedding

208.3, 60, 88.2

color rep

DecoderEncoder

208.3, 60, 88.2
color embedding
color rep
dark<s> blue
x1 x20 x11
h1 h2 h3
dark blue </s>
Color descriptor of Monroe et al. (2016)
### Colors in context (Monroe et al. 2017)

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ] blue</td>
<td>The darker blue one</td>
</tr>
<tr>
<td>[ ] teal</td>
<td>teal not the two that are more green</td>
</tr>
<tr>
<td>[ ] dull pink</td>
<td>dull pink not the super bright one</td>
</tr>
<tr>
<td>[ ] greens</td>
<td>not any of the regular greens</td>
</tr>
<tr>
<td>[ ] Purple</td>
<td>Purple</td>
</tr>
<tr>
<td>[ ] blue</td>
<td>blue</td>
</tr>
</tbody>
</table>

**Table:** Examples from the Colors in Context corpus from the Stanford Computation & Cognition Lab
Colors in context (Monroe et al. 2017)
Related ideas and tasks

• The preceding can be seen as a special case of *image captioning*, which has been revolutionized by neural methods in recent years (Karpathy & Fei-Fei 2015; Vinyals et al. 2015).

• The Encoder part of captioning models is likely to be more involved than the above, but the basic structure is the same.

• Mao et al. (2016) and Vedantam et al. (2017) explore variants of the image captioning task that are like the ‘colors in context’ task above.

• Visual Question Answering is a more structured variant of the problem in which an image and a question text are the inputs and the goal is to produce grounded answers.
Listeners: From language to the world

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Listeners: From language to the world
4. Grounded chat bots
5. Reasoning about other minds
6. A few other grounding ideas
### Color interpreter: Task formulation and data

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Color Context" /></td>
<td><img src="image" alt="Utterance" /></td>
</tr>
<tr>
<td>blue</td>
<td>The darker blue one</td>
</tr>
<tr>
<td>teal</td>
<td>teal not the two that are more green</td>
</tr>
<tr>
<td>dull pink</td>
<td>dull pink not the super bright one</td>
</tr>
<tr>
<td>not any of the regular greens</td>
<td>not any of the regular greens</td>
</tr>
<tr>
<td>Purple</td>
<td>Purple</td>
</tr>
<tr>
<td>blue</td>
<td>blue</td>
</tr>
</tbody>
</table>

**Table**: Examples from the Colors in Context corpus from the Stanford Computation & Cognition Lab
A neural listener model

Encoder

Decoder

\[ x_{37} \rightarrow h_1 \rightarrow h_2 \rightarrow x_{11} \]

light → blue
A neural listener model

Diagram:

Encoder

- $h_1$
- $x_{37}$
- light

Decoder

- $h_2$
- $x_{11}$
- blue

Connection:

$h_1$ to $h_2$

$(\mu, \Sigma)$
A neural listener model

Encoder

Decoder

$h_1 \xrightarrow{} h_2 \xrightarrow{} (\mu, \Sigma)$

$x_{37} \leftarrow \text{light}

x_{11} \leftarrow \text{blue}

\begin{align*}
c_1 & \quad c_2 & \quad c_T
\end{align*}
A neural listener model

Encoder

Decoder

Encoder: $h_1$, $h_2$, $x_{37}$, $x_{11}$, light, blue

Decoder: $f_1$, $f_2$, $f_T$, $c_1$, $c_2$, $c_T$, $(\mu, \Sigma)$

Fourier transform
### A neural listener model

**Encoder**

- $h_1$
- $x_{37}$: light
- $h_2$
- $x_{11}$: blue

**Decoder**

- $(\mu, \Sigma)$
- $s_1$
- $s_2$
- $s_T$

**Score Function**

$$ \text{score}(f_i) = -(f_i - \mu)^T \Sigma (f_i - \mu) $$

**Fourier Transform**

- $f_1$
- $f_2$
- $f_T$
- $c_1$
- $c_2$
- $c_T$
A neural listener model

Encoder

h₁

x₁₁

light

blue

x₃₇

Decoder

softmax(s₁, s₂, s₃)

(μ, Σ)

score(fᵢ) = -(fᵢ - μ)ᵀΣ(fᵢ - μ)

Fourier transform

f₁

f₂

fₜ

c₁

c₂

cₜ
Other ideas and datasets

- **NLU classifiers** are very simple listeners: they consume language and make an inference in a structured space.

- **Semantic parsers** are very complex listeners: they consume language, construct rich latent representations, and predict into structured output spaces.

- **Scene generation** is the task of mapping language to structured representations of visual scenes (Seversky & Yin 2006; Chang et al. 2014, 2015).

- Young et al. (2014) seek to learn visual denotations for linguistic expressions.
Grounded chat bots

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Listeners: From language to the world
4. Grounded chat bots
5. Reasoning about other minds
6. A few other grounding ideas
Basic neural chatbot
FAIR negotiation dataset

5,808 dialogues grounded in 2,236 unique scenarios.

Figure 1: A dialogue in our Mechanical Turk interface, which we used to collect a negotiation dataset.

From Lewis et al. 2017; see also Yarats & Lewis 2018
FAIR negotiation dataset

Perspective of YOU

1. 1 0 4 2 1 2 # (1 book, worth 0; 4 hats, worth 2, 1 ball, worth 2)
2. YOU: i would like 4 hats and you can have the rest <eos>
   THEM: deal <eos>
   YOU: <selection>
3. item0=0 item1=4 item2=0
4. <eos>
5. reward=8
6. agree
7. 1 4 4 1 1 2
FAIR negotiation dataset

Perspective of THEM

1. 1 4 4 1 1 2  # (1 book, worth 4; 4 hats, worth 1, 1 ball, worth 2)
2. THEM: i would like 4 hats and you can have the rest <eos>
   YOU: deal <eos>
   THEM: <selection>
3. item0=1 item1=0 item2=1
4. <eos>
5. reward=6
6. agree
7. 1 0 4 2 1 2
FAIR negotiation agents
Goal-based training

Goal encoder GRU

Agent A reads

Dialogue encoder GRU

Agent A writes
Agent B writes
Agent A reads

Output encoder GRU

Agent A reward

item0=1
item1=4
item2=1
Witem=0
Witem=1
Witem=2

h^g

h^s

attention vector

x_1 h^g x_2 h^g x_3 h^g x_4 h^g

h_1 h_2 h_3 h_3

four hats </s> deal <select>

Witem=0
Witem=1
Witem=2
Decoding through rollouts

For Lewis et al. 2017, figure 4
Aside: An amusing media narrative

**Lewis et al. (2017)**

“During reinforcement learning, an agent $A$ attempts to improve its parameters from conversations with another agent $B$. While the other agent $B$ could be a human, in our experiments we used our fixed supervised model that was trained to imitate humans. The second model is fixed as we found that updating the parameters of both agents led to divergence from human language.”
Aside: An amusing media narrative

**FAIR blog post [link]**

“The second model is fixed, because the researchers found that updating the parameters of both agents led to divergence from human language as the agents developed their own language for negotiating.”
Aside: An amusing media narrative

Newsweek [link]

“The bots ran afoul of their Facebook overlords when they started to make up their own language to do things faster, not unlike the way football players have shorthand names for certain plays instead of taking the time in the huddle to describe where everyone should run. It’s not unusual for bots to make up a lingo that humans can’t comprehend, though it does stir worries that these things might gossip about us behind our back. Facebook altered the code to make the bots stick to plain English.”
Aside: An amusing media narrative

Tech Times [link]

“Facebook was forced to shut down one of its artificial intelligence systems after researchers discovered that it had started communicating in a language that they could not understand.
Aside: An amusing media narrative

Tech Times [link]

“Facebook was forced to shut down one of its artificial intelligence systems after researchers discovered that it had started communicating in a language that they could not understand.

“The incident evokes images of the rise of Skynet in the iconic Terminator series. Perhaps Tesla CEO Elon Musk is right about AI being the ‘biggest risk we face.’”
Other task-oriented dialogue datasets

- Edinburgh Map Corpus
  [http://groups.inf.ed.ac.uk/maptask/](http://groups.inf.ed.ac.uk/maptask/)

- TRIPS

- TRAINS

- Cards
  [http://CardsCorpus.christopherpotts.net/](http://CardsCorpus.christopherpotts.net/)

- SCARE
  [http://slate.cse.ohio-state.edu/quake-corpora/scare/](http://slate.cse.ohio-state.edu/quake-corpora/scare/)

- The Carnegie Mellon Communicator Corpus
  [http://www.speech.cs.cmu.edu/Communicator/](http://www.speech.cs.cmu.edu/Communicator/)
Reasoning about other minds

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Listeners: From language to the world
4. Grounded chat bots
5. Reasoning about other minds
6. A few other grounding ideas
Pragmatic reasoning à la Grice (1975)
Pragmatic reasoning à la Grice (1975)
Pragmatic reasoning à la Grice (1975)
Pragmatic reasoning à la Grice (1975)

The blue one, please!
Pragmatic reasoning à la Grice (1975)

My listener knows I’m cooperative in the Gricean sense.

The blue one, please!
Pragmatic reasoning à la Grice (1975)

My listener knows I'm cooperative in the Gricean sense.

So they will be able to work out that I mean the unmarked blue.

The blue one, please!
Pragmatic reasoning à la Grice (1975)

My listener knows I’m cooperative in the Gricean sense.

The blue one, please!

The speaker’s utterance seems ambiguous or under-informative.

So they will be able to work out that I mean the unmarked blue.
Pragmatic reasoning à la Grice (1975)

My listener knows I’m cooperative in the Gricean sense.

The speaker’s utterance seems ambiguous or under-informative.

The blue one, please!

So they will be able to work out that I mean the unmarked blue.

But I’m assuming the speaker is cooperative in the Gricean sense!
Pragmatic reasoning à la Grice (1975)

My listener knows I'm cooperative in the Gricean sense.

So they will be able to work out that I mean the unmarked blue.

The speaker's utterance seems ambiguous or under-informative.

But I'm assuming the speaker is cooperative in the Gricean sense!

Ah, but if I assume they would have picked a marked form like "baby blue" if it were true, then I can work out what they want!
Pragmatic reasoning à la Grice (1975)

The speaker’s utterance seems ambiguous or under-informative. But I’m assuming the speaker is cooperative in the Gricean sense!

Ah, but if I assume they would have picked a marked form like “baby blue” if it were true, then I can work out what they want!

My listener knows I’m cooperative in the Gricean sense.

So they will be able to work out that I mean the unmarked blue.

The blue one, please!

The speaker’s utterance seems ambiguous or under-informative. But I’m assuming the speaker is cooperative in the Gricean sense!

Ah, but if I assume they would have picked a marked form like “baby blue” if it were true, then I can work out what they want!
Pragmatic reasoning à la Grice (1975)

The blue one, please!
The Rational Speech Acts Model

(Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Goodman & Frank 2016)
The Rational Speech Acts Model

Literal listener

\[ l_0(w \mid \text{msg}, \text{Lex}) \propto \text{Lex}(\text{msg}, w)P(w) \]

(Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Goodman & Frank 2016)
The Rational Speech Acts Model

Pragmatic speaker

\[ s_1(msg \mid w, Lex) \propto \exp \lambda (\log l_0(w \mid msg, Lex) - C(msg)) \]

Literal listener

\[ l_0(w \mid msg, Lex) \propto Lex(msg, w)P(w) \]

(Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Goodman & Frank 2016)
The Rational Speech Acts Model

**Pragmatic listener**

\[ l_1(w \mid \text{msg, \text{Lex}}) \propto s_1(\text{msg} \mid w, \text{Lex})P(w) \]

**Pragmatic speaker**

\[ s_1(\text{msg} \mid w, \text{Lex}) \propto \exp \lambda (\log l_0(w \mid \text{msg, \text{Lex}}) - C(\text{msg})) \]

**Literal listener**

\[ l_0(w \mid \text{msg, \text{Lex}}) \propto \text{Lex}(\text{msg}, w)P(w) \]

(Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Goodman & Frank 2016)
The Rational Speech Acts Model

Pragmatic listener

\[ l_1(w \mid msg, Lex) = \text{pragmatic speaker} \times \text{state prior} \]

Pragmatic speaker

\[ s_1(msg \mid w, Lex) = \text{literal listener} - \text{message costs} \]

Literal listener

\[ l_0(w \mid msg, Lex) = \text{lexicon} \times \text{state prior} \]

(Frank & Goodman 2012; Goodman & Stuhlmüller 2013; Goodman & Frank 2016)
RSA listener example

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>beard</td>
<td>T</td>
<td>F</td>
</tr>
<tr>
<td>glasses</td>
<td>T</td>
<td>T</td>
</tr>
</tbody>
</table>

$\begin{align*}
  l_1 \\
  s_1 \\
  l_0 \\
  \text{Lex}
\end{align*}$
RSA listener example

<table>
<thead>
<tr>
<th>词汇</th>
<th>讲者</th>
<th>听众</th>
</tr>
</thead>
<tbody>
<tr>
<td>beard</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>glasses</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
I_1 & \quad s_1 \\
I_0 & \quad \text{Lex}
\end{align*}
\]
### RSA listener example

<table>
<thead>
<tr>
<th>beard</th>
<th>glasses</th>
</tr>
</thead>
<tbody>
<tr>
<td>.67</td>
<td>.33</td>
</tr>
</tbody>
</table>

![RSA listener example](image-url)
RSA listener example

<table>
<thead>
<tr>
<th>beard</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>glasses</td>
<td>.25</td>
<td>.75</td>
</tr>
</tbody>
</table>

\( l_1 \)
\( s_1 \)
\( l_0 \)
\( \text{Lex} \)
Limitations

- Hand-specified lexicon
- Reasoning about all possible utterances?

\[ s_1(msg \mid w, \text{Lex}) = \frac{l_0(w \mid msg, \text{Lex})}{\sum_{msg'} l_0(w \mid msg', \text{Lex})} \]

- High-bias model; few chances to learn from data

<table>
<thead>
<tr>
<th></th>
<th>beard</th>
<th>glasses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>.25</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>.75</td>
</tr>
</tbody>
</table>
### Colors in context (Monroe et al. 2017)

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Blue" /></td>
<td>blue</td>
</tr>
<tr>
<td><img src="image" alt="Blue" /></td>
<td>The darker blue one</td>
</tr>
<tr>
<td><img src="image" alt="Teal" /></td>
<td>teal not the two that are more green</td>
</tr>
<tr>
<td><img src="image" alt="Pink" /></td>
<td>dull pink not the super bright one</td>
</tr>
<tr>
<td><img src="image" alt="Green" /></td>
<td>not any of the regular greens</td>
</tr>
<tr>
<td><img src="image" alt="Purple" /></td>
<td>Purple</td>
</tr>
<tr>
<td><img src="image" alt="Blue" /></td>
<td>blue</td>
</tr>
</tbody>
</table>

**Table:** Examples from the Colors in Context corpus from the Stanford Computation & Cognition Lab
Literal neural speaker $S_0$
Neural literal listener $\mathcal{L}_0$
Neural pragmatic agents

Neural pragmatic speaker (Andreas & Klein 2016)

\[ S_1(msg | c, C; \theta) = \frac{\mathcal{L}_0(c | msg, C; \theta)}{\sum_{msg' \in X} \mathcal{L}_0(c | msg', C; \theta)} \]

where \( X \) is a sample from \( S_0(msg | c, C; \theta) \) such that \( msg^* \in X \).

Neural pragmatic listener

\[ \mathcal{L}_1(c | msg, C; \theta) \propto S_1(msg | c, C; \theta) \]

Blended neural pragmatic listener

Weighted combination of \( \mathcal{L}_0 \) and \( \mathcal{L}_1 \).
Pragmatic image captioning

Mao et al. (2016); Vedantam et al. (2017): Captions that are true and distinguish their images from related ones.

$S_0$ caption: the dog is brown
$S_1$ caption: the head of a dog

Reasoning about all possible utterances/captions?

(Cohn-Gordon et al. 2018, 2019)
Pragmatic image captioning

Mao et al. (2016); Vedantam et al. (2017): Captions that are true and distinguish their images from related ones.

 Reasoning about all possible utterances/captions?  
⇒ Sample from $S_0$

(Cohn-Gordon et al. 2018, 2019)
Pragmatic image captioning

Mao et al. (2016); Vedantam et al. (2017): Captions that are true and distinguish their images from related ones.

*S₀ caption: the dog is brown
*S₁ caption: the head of a dog

Reasoning about all possible utterances/captions?

⇒ Full RSA reasoning about characters

(Cohn-Gordon et al. 2018, 2019)
Other related work

- Golland et al. (2010): Recursive speaker/listener reasoning as part of interpreting complex utterances compositionally, with grounding in a simple visual world.
- Tellex et al.’s (2014) Inverse Semantics: Robot utterances are scored by models similar to RSA’s pragmatic speakers.
- Wang et al. (2016): Pragmatic reasoning helps in online learning of semantic parsers.
- Monroe & Potts (2015): “RSA as a hidden activation function”
- Fried et al. (2018): Sequential instruction following with pragmatic reasoning.
Other relevant datasets

- The TUNA Reference Corpus
  https://www.abdn.ac.uk/ncs/departments/computing-science/corpus-496.php

- SCONE: Sequential CONtext-dependent Execution
  https://nlp.stanford.edu/projects/scone/

- Crowdsource your own (Hawkins 2015)!
  https://github.com/hawkrobe/MWERT
A few other grounding ideas

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Listeners: From language to the world
4. Grounded chat bots
5. Reasoning about other minds
6. A few other grounding ideas
Modeling users for sarcasm detection

(SARC: Khodak et al. 2017; Kolchinski & Potts 2018)
NLU in social graphs with Probabilistic Soft Logic

I actually kind of liked it.

Bla bla … sentiment … bla bla bla … networks …

Dude, that was even more boring than his gray shirt, eh?!

Yeah right. Great talk… He didn’t even talk about deep learning.

(PSL: https://psl.linqs.org; West et al. 2014)
NLU in social graphs with Probabilistic Soft Logic

Yeah right. Great talk... He didn’t even talk about deep learning.

Social balance theory

“The friend of my enemy is my enemy”

(PSL: https://psl.linqs.org; West et al. 2014)
PLOW: Webpage structure as context

1. Learning rules of the form ‘If A, then B, else C’ is a challenge because the latent variable A is generally not observed. Rather, one sees only B or C.

2. In an interactive, instructional setting, one needn’t rely entirely on abduction or probabilistic inference: users generally state the needed rules during their interactions.

3. The user’s actions ground the parsed language.

4. The DOM structure grounds the user’s indexicals:
   - Put the name here. (user clicks on the DOM element)
   - This is the ISBN number. (user highlights some text)
   - Find another tab. (user has selected a tab)

(Allen et al. 2007)
Decision-theoretic agents

Both players must find the ace of spades. DialogBot:

(Vogel et al. 2013a,b)
Decision-theoretic agents

Baby DialogBots (a few hours of policy exploration)

(Vogel et al. 2013a,b)
Decision-theoretic agents

Grown-up DialogBots (a week of policy exploration)

(Vogel et al. 2013a,b)
Frontiers

- Deeper integration with devices and the environment.
- More sophisticated reasoning about other agents and their goals.
- Better tracking of full dialogue history; improved discourse coherence.
- Approximate state representations to address very pressing scalability issues.
References


