Grounded language understanding

Christopher Potts

Stanford Linguistics

CS 224U: Natural language understanding
May 4 and 6
Overview

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. A few other grounding ideas
Associated materials

1. Code
   a. Notebook: colors_overview.ipynb
   b. Homework and bake-off: hw_colors.ipynb

2. Core reading: Monroe et al. 2017

3. Auxiliary readings: Golland et al. 2010; Lewis et al. 2017; Andreas and Klein 2016; Tellex et al. 2014; Vogel et al. 2013a
HAL

- 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

Graphics

HAL

Jurassic Park (1993)

Slide idea from Andrew McCallum
HAL

Chess

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.]
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.”
Levinson’s (2000) analogy

Figure 0.1
Rembrandt sketch
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 […].
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.”
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)
- 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

“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)
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.
Routine pragmatic enrichment

how big is the contextually restricted domain of students?

what’s the additional contextual restriction?

false for most students?

Many students met with me yesterday.

what’s the time of utterance?

but perhaps many met with the speaker at other times?

who’s the speaker?
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.”
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

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 and Goodman 2014
Consequences for NLU

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

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

3. 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.

4. 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. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. 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>

McMahan and Stone 2015
Color describer: Training with *teacher forcing*
Color describer: Training with *teacher forcing*

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

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

Decoder

- Dark
  - h₁
  - x₁
  - <s>

Decoder

- Blue
  - h₂
  - x₃₇
  - light
  - blue

Decoder

- Green
  - h₃
  - x₁₁
Color describer: Training with *teacher forcing*

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

**Decoder**
- **h_1**: dark
- **h_2**: blue
- **h_3**: green
- **x_1**: light
- **x_{37}**: light
- **x_{11}**: blue
- **<s>**
- **</s>**
Color describer: Prediction

Encoder

color embedding

208.3, 60, 88.2

color rep

Decoder
Color describer: Prediction

Encoder

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

Decoder

<s>
Color describer: Prediction
Color describer: Prediction
Color describer: Prediction

Encoder

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

Decoder

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

Encoder

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

Decoder

- Dark
- Encoder
- Decoder
- Color embedding
- Color rep
- 208.3, 60, 88.2
- x₁
- h₁
- h₂
- x₂₀
- <s>
- Dark
Color describer: Prediction
Color describer: Prediction

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

**Decoder**
- dark
- blue
- $x_1$, $h_1$, $x_{20}$, $h_2$, $x_{11}$, $h_3$

- dark $<$s$>$ blue

- decoder $\rightarrow$ encoder

- hidden states $h_1$, $h_2$, $h_3$

- input $x_1$, $x_{20}$, $x_{11}$
Color describer: Prediction

Encoder

Decoder

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

- **dark**
- **blue**
- **</s>**

- **x₁**
- **x₂₀**
- **x₁₁**

**Diagram:**
- Encoder and Decoder blocks connected with arrows,
- Color representation and embedding are linked,
- Embedding color values are shown.
**Miscellaneous design choices**

- The Encoder and Decoder could have more hidden layers. We would expect the layer counts to match to facilitate the hand-off between Encoder and Decoder, though pooling or copying might work too.

- It seems very common at present for researchers to tie the embedding and classifier parameters (Press and Wolf 2017).

- During training, one might drop teacher forcing a small percentage of the time to encourage the model to explore.

- Some Encoder states could be appended to some or all of the Decoder inputs. (See the next slide and the assignment.)
Color descriptor of Monroe et al. (2016)

Encoder
- Color rep
- Fourier transform
- HSV

Decoder
- w₂
- h₁
- x₁
- <s>
- light
- blue
- w₃
- h₂
- x₃₇
- w₄
- h₃
- x₁₁
Related tasks

Non-linguistic representation $\Rightarrow$ Language

- Image captioning
- Scene description
- Visual Question Answering
  (Image + Question-text $\Rightarrow$ Answer-text)
- Instruction giving (State $\Rightarrow$ Language)
- ...
Overview of the assignment and bake-off

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. A few other grounding ideas
## Color descriptions in context

<table>
<thead>
<tr>
<th>Color</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green</td>
<td>green</td>
</tr>
<tr>
<td>Purple</td>
<td>purple</td>
</tr>
<tr>
<td>Grape</td>
<td>grape</td>
</tr>
<tr>
<td>Turquoise</td>
<td>turquoise</td>
</tr>
<tr>
<td>Moss green</td>
<td>moss green</td>
</tr>
<tr>
<td>Pinkish purple</td>
<td>pinkish purple</td>
</tr>
<tr>
<td>Light blue grey</td>
<td>light blue grey</td>
</tr>
<tr>
<td>Robin’s egg blue</td>
<td>robin’s egg blue</td>
</tr>
<tr>
<td>British racing green</td>
<td>british racing green</td>
</tr>
<tr>
<td>Baby puke green</td>
<td>baby puke green</td>
</tr>
</tbody>
</table>

McMahan and Stone 2015
## Color descriptions in context

<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="blue.png" alt="Blue" /></td>
<td>blue</td>
</tr>
<tr>
<td><img src="darkblue.png" alt="Dark Blue" /></td>
<td>The darker blue one</td>
</tr>
<tr>
<td><img src="teal.png" alt="Teal" /></td>
<td>teal not the two that are more green</td>
</tr>
<tr>
<td><img src="dullpink.png" alt="Dull Pink" /></td>
<td>dull pink not the super bright one</td>
</tr>
<tr>
<td><img src="notreggreens.png" alt="Not Regular Greens" /></td>
<td>not any of the regular greens</td>
</tr>
<tr>
<td><img src="purple.png" alt="Purple" /></td>
<td>Purple</td>
</tr>
<tr>
<td><img src="blue.png" alt="Blue" /></td>
<td>blue</td>
</tr>
</tbody>
</table>

Stanford Colors in Context corpus (Monroe et al. 2017)
Colors in context (Monroe et al. 2017)
Data overview

```python
[1]: from colors import ColorsCorpusReader
    import os

[2]: COLORS_SRC_FILENAME = os.path.join("data", "colors", "filteredCorpus.csv")

[3]: corpus = ColorsCorpusReader(COLORS_SRC_FILENAME, normalize_colors=True)

[4]: examples = list(corpus.read())

[5]: len(examples)

[5]: 46994
```

More details: colors_overview.ipynb
## Data overview

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>[6]:</td>
<td><code>examples[0].display()</code></td>
</tr>
<tr>
<td></td>
<td>The darker blue one</td>
</tr>
<tr>
<td>[7]:</td>
<td><code>examples[0].colors</code></td>
</tr>
<tr>
<td>[7]:</td>
<td><code>[[0.7861111111111111, 0.5, 0.87],</code></td>
</tr>
<tr>
<td></td>
<td><code>[0.6888888888888889, 0.5, 0.92],</code></td>
</tr>
<tr>
<td></td>
<td><code>[0.6277777777777778, 0.5, 0.81]]</code></td>
</tr>
<tr>
<td>[8]:</td>
<td><code>examples[0].contents</code></td>
</tr>
<tr>
<td>[8]:</td>
<td>'The darker blue one'</td>
</tr>
</tbody>
</table>

More details: `colors_overview.ipynb`
**Data overview**

```
[9]: print("Condition type:", examples[1].condition)

examples[1].display()

Condition type: far
purple
```

More details: colors_overview.ipynb
Data overview

```python
[10]: print("Condition type:", examples[3].condition)
    examples[3].display()

Condition type: split
lime

More details: colors_overview.ipynb
Data overview

[11]: print("Condition type:", examples[2].condition)

examples[2].display()

Condition type: close
Medium pink ### the medium dark one

More details: colors_overview.ipynb
Task-oriented evaluation

Predictions
For a given context $c$, let $C$ be all of its permutations. Then a speaker model $P_S$ predicts:

$$c^* = \text{argmax}_{c \in C} P_S(msg \mid c)$$
Task-oriented evaluation

Predictions
For a given context $c$, let $C$ be all of its permutations. Then a speaker model $P_S$ predicts:

$$c^* = \arg\max_{c \in C} P_S(msg \mid c)$$

Accuracy
A speaker model $P_S$ is correct in its prediction about $c$ iff $c^*[-1]$ is the target.
Task-oriented evaluation

Predictions
For a given context $c$, let $C$ be all of its permutations. Then a speaker model $P_S$ predicts:

$$c^* = \arg\max_{c \in C} P_S(msg \mid c)$$

Example
$\text{msg} = \text{“blue”}$

$c = \text{[color1, color2, color3]}$
Task-oriented evaluation

Predictions
For a given context $c$, let $C$ be all of its permutations. Then a speaker model $P_S$ predicts:

$$c^* = \arg\max_{c \in C} P_S(msg | c)$$

Example

$\text{msg} = \text{“blue”}$

$\text{c} = \boxed{\text{G}} \boxed{\text{R}} \boxed{\text{H}}$
Question 1: Improve the tokenizer

```python
[ ]: import utils
    from utils import START_SYMBOL, END_SYMBOL, UNK_SYMBOL

[ ]: def tokenize_example(s):
      # Improve me!
      return [START_SYMBOL] + s.split() + [END_SYMBOL]
```
Question 2: Improve the color representations

```python
[ ]: def represent_color_context(colors):
    # Improve me!
    return [represent_color(color) for color in colors]

def represent_color(color):
    # Improve me!
    return color
```
Question 3: GloVe embeddings

```python
[ ]: def create_glove_embedding(vocab, glove_base_filename='glove.6B.50d.txt):

    # Use `utils.glove2dict` to read in the GloVe file:
    ###### YOUR CODE HERE

    # Use `utils.create_pretrained_embedding` to create the embedding.
    # This function will, by default, ensure that START_TOKEN, END_TOKEN, and UNK_TOKEN are included in the embedding.
    ###### YOUR CODE HERE

    # Be sure to return the embedding you create as well as the vocabulary returned by `utils.create_pretrained_embedding`,
    # which is likely to have been modified from the input `vocab`.

    ###### YOUR CODE HERE
```
Question 4: Color context

![Diagram of a neural network model with encoder and decoder components. The encoder includes nodes labeled 'distractor', 'distractor', and 'target'. The decoder includes nodes labeled 'w2', 'w3', 'w4', 'h1', 'h2', 'h3', 'x1', 'x37', and 'x11'. Words include '<s>', 'light', and 'blue'.]
Question 4: Color context

1. Modify Decoder so that the input vector to the model at each timestep is not just a token representation $x$ but the concatenation of $x$ with the representation of the target color.

2. Modify EncoderDecoder to extract the target colors and feed them to the decoder.

3. Modify ContextualColorDescriber so that it uses your new Decoder and EncoderDecoder.

   Use the toy dataset generator for development!
Original system and bake-off

Our expectation for how you’ll work:

1. Iteratively improve answers to the assignment questions.
2. Perhaps extend your modified Encoder/Decoder classes to do interesting new things.
3. Any data you can find is fine for your development work.
4. The bake-off will involve a new test set that has never been released anywhere before:
   - Same kinds of color context as in the released corpus.
   - One-off games rather than iterated.
   - All items listener-validated.
Listeners: From language to the world

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. 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>blue</td>
<td>blue</td>
</tr>
<tr>
<td>The darker blue one</td>
<td>teal not the two that are more green</td>
</tr>
<tr>
<td>dull pink not the super bright one</td>
<td>not any of the regular greens</td>
</tr>
<tr>
<td>Purple</td>
<td>blue</td>
</tr>
</tbody>
</table>

Stanford Colors in Context corpus (Monroe et al. 2017)
A neural listener model

Encoder

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

\[ x_{11} \rightarrow \text{blue} \]

\[ \text{light} \]

Decoder

\[ (\mu, \Sigma) \]

\[ s_1, s_2, s_T \]

\[ \text{softmax}(s_1, s_2, s_3) \]

score\( (f_i) = -(f_i - \mu)^T \Sigma (f_i - \mu) \)

Fourier transform

\[ f_1, f_2, f_T \]

\[ c_1, c_2, c_T \]
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 and Yin 2006; Chang et al. 2014, 2015).

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

- Mei et al. (2015) develop essentially a seq2seq version of the above model: given a linguistic input, they predict action sequences. (Kai Sheng Tai did his 2015 CS224u project on this, working at the same time as Mei et al.!) 

- Suhr et al. (2019): Released the CerealBar data and game engine for learning to execute instructions.
Reasoning about other minds

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. **Reasoning about other minds**
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. A few other grounding ideas
Discriminative image labeling

Mao et al. 2016
Discriminative image labeling

Mao et al. 2016
Discriminative image labeling

Mao et al. 2016
Discriminative image labeling

A little dog jumping and catching a frisbee

A big dog running

Mao et al. 2016
Discriminative image captioning

Vedantam et al. 2017; Cohn-Gordon et al. 2018
Discriminative image captioning

Vedantam et al. 2017; Cohn-Gordon et al. 2018
Discriminative image captioning

Vedantam et al. 2017; Cohn-Gordon et al. 2018
Summarization

Tennis champion Serena Williams wobbled into the Third Round of the Australian Open on Thursday.

Serena Williams advances to Australian Open Third Round.

Ongoing work with Hanson Lu and Reuben Cohn-Gordon
Tennis champion Serena Williams wobbled into the Third Round of the Australian Open on Thursday.

Serena Williams advances to Australian Open Third Round.

Sports Champion advances in tournament.

Ongoing work with Hanson Lu and Reuben Cohn-Gordon
Summarization

Tennis champion Serena Williams wobbled into the Third Round of the Australian Open on Thursday.

Serena Williams advances to Australian Open Third Round.

Sports Champion advances in tournament.

Williams wobbled on Thursday.

Ongoing work with Hanson Lu and Reuben Cohn-Gordon
Summarization

Tennis champion Serena Williams wobbled into the Third Round of the Australian Open on Thursday.

Olympic Gold Medalist Venus Williams advanced to the US Open Semi-Finals on Friday.

Serena Williams advances to Australian Open Third Round.

Sports Champion advances in tournament.

Williams wobbled on Thursday.

Ongoing work with Hanson Lu and Reuben Cohn-Gordon
Summarization

Tennis champion Serena Williams wobbled into the Third Round of the Australian Open on Thursday.

Olympic Gold Medalist Venus Williams advanced to the US Open Semi-Finals on Friday.

Golfer Lydia Ko eliminated from British Open after finishing 12 over par.

Serena Williams advances to Australian Open Third Round.

Sports Champion advances in tournament.

Williams wobbled on Thursday.

Ongoing work with Hanson Lu and Reuben Cohn-Gordon
Machine translation

She chopped up the tree. \rightarrow Elle coupa l’arbre.

She chopped down the tree. \rightarrow Elle a abattu l’arbre.
**Machine translation**

- She chopped up the tree. → Elle coupa l’arbre.
- She chopped down the tree. → Elle a abattu l’arbre.

Cohn-Gordon and Goodman 2019
Machine translation

She chopped up the tree. → Elle coupa l’arbre.

Elle coupa l’arbre. → Elle coupé l’arbre.

Elle coupé l’arbre. → Elle a abattu l’arbre.

She chopped down the tree. → Elle a abattu l’arbre.

Cohn-Gordon and Goodman 2019
Generating and following instructions

Behavior

(a) Base Speaker

walk forward four times

go forward four segments to the intersection with the bare concrete hall

(b) Base Listener

Instruction

walk along the blue carpet and you pass two objects

Rational Listener

Fried et al. 2018
Collaborative problem solving

Tellex et al. 2014
Collaborative problem solving

Help me!

Tellex et al. 2014
Collaborative problem solving

Hand me the leg

Tellex et al. 2014
Collaborative problem solving

Hand me the white leg on the table

Tellex et al. 2014
Optical character recognition
The Rational Speech Acts model

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. A few other grounding ideas
Origin story

- Rosenberg and Cohen (1964): early Bayesian model of production and comprehension
- Lewis (1969): signaling systems
- Rabin (1990): recursive strategic signaling
- Camerer et al. (2004): cognitive hierarchy models for games of conflict and coordination
- Golland et al. (2010): pragmatic listeners and probabilistic compositionality
- Frank and Goodman (2012): very sophisticated pragmatic agents and a new Bayesian foundation See also Goodman and Stuhlmüller 2013.
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.

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 listeners
Pragmatic listeners

Literal listener

\[ L_{\text{lit}}(state \mid msg) = \frac{[msg, state]P(state)}{\sum_{state'}[msg, state']P(state')} \]
Pragmatic listeners

Pragmatic speaker

\[ S_{\text{prag}}(msg \mid state) = \frac{\exp(\alpha (\log L_{\text{lit}}(state \mid msg) - C(msg)))}{\sum_{msg'} \exp(\alpha (\log L_{\text{lit}}(state \mid msg') - C(msg')))} \]

Literal listener

\[ L_{\text{lit}}(state \mid msg) = \frac{\llbracket msg, state \rrbracket P(state)}{\sum_{state'} \llbracket msg, state' \rrbracket P(state')} \]
Pragmatic listeners

Pragmatic listener

\[ L_{\text{prag}}(\text{state} \mid \text{msg}) = \frac{S_{\text{prag}}(\text{msg} \mid \text{state})P(\text{state})}{\sum_{\text{state}'} S_{\text{prag}}(\text{msg} \mid \text{state}')}P(\text{state}')} \]

Pragmatic speaker

\[ S_{\text{prag}}(\text{msg} \mid \text{state}) = \frac{\exp(\alpha (\log L_{\text{lit}}(\text{state} \mid \text{msg}) - C(\text{msg})))}{\sum_{\text{msg}'} \exp(\alpha (\log L_{\text{lit}}(\text{state} \mid \text{msg}') - C(\text{msg}')))} \]

Literal listener

\[ L_{\text{lit}}(\text{state} \mid \text{msg}) = \frac{[\text{msg, state}]P(\text{state})}{\sum_{\text{state}'} [\text{msg, state}]P(\text{state}')} \]
Pragmatic listeners

Pragmatic listener

\[ L_{\text{prag}}(state \mid msg) = \text{pragmatic speaker} \times \text{state prior} \]

Pragmatic speaker

\[ S_{\text{prag}}(msg \mid state) = \text{literal listener} - \text{message costs} \]

Literal listener

\[ L_{\text{lit}}(state \mid msg) = \text{lexicon} \times \text{state prior} \]
## A simple example

<table>
<thead>
<tr>
<th>Item</th>
<th>$L_{\text{lit}}$</th>
<th>$S_{\text{prag}}$</th>
<th>$L_{\text{prag}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>beard</td>
<td>1 0 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>glasses</td>
<td>1 1 0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tie</td>
<td>0 1 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A simple example

<table>
<thead>
<tr>
<th></th>
<th>$L_{\text{prag}}$</th>
<th>$S_{\text{prag}}$</th>
<th>$L_{\text{lit}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>beard</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>glasses</td>
<td>.5</td>
<td>.5</td>
<td>0</td>
</tr>
<tr>
<td>tie</td>
<td>0</td>
<td>.5</td>
<td>.5</td>
</tr>
</tbody>
</table>
A simple example

<table>
<thead>
<tr>
<th>beard</th>
<th>glasses</th>
<th>tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>.67</td>
<td>.33</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

$L_{\text{prag}}$ $S_{\text{prag}}$ $L_{\text{lit}}$ $[\cdot]$
A simple example

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>beard</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>glasses</td>
<td>0.25</td>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>tie</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$L_{prag}$
$S_{prag}$
$L_{lit}$
$[\cdot\cdot]$
Pragmatic speakers
Pragmatic speakers

Literal speaker

\[ S_{\text{lit}}(msg \mid state) = \frac{\exp (\alpha (\log [msg, state] - C(msg)))}{\sum_{msg'} \exp (\alpha (\log [msg', state] - C(msg')))} \]
Pragmatic speakers

Pragmatic listener

\[
L_{\text{prag}}(\text{state} \mid \text{msg}) = \frac{S_{\text{lit}}(\text{msg} \mid \text{state})P(\text{state})}{\sum_{\text{state}'} S_{\text{lit}}(\text{msg} \mid \text{state}')P(\text{state}')}
\]

Literal speaker

\[
S_{\text{lit}}(\text{msg} \mid \text{state}) = \frac{\exp (\alpha (\log \| \text{msg}, \text{state} \| - C(\text{msg})))}{\sum_{\text{msg}'} \exp (\alpha (\log \| \text{msg}', \text{state} \| - C(\text{msg}')))}
\]
Pragmatic speakers

Pragmatic speaker

\[ S_{\text{prag}}(msg \mid state) = \frac{\exp(\alpha (\log L_{\text{prag}}(state \mid msg) - C(msg)))}{\sum_{msg'} \exp(\alpha (\log L_{\text{prag}}(state \mid msg') - C(msg')))} \]

Pragmatic listener

\[ L_{\text{prag}}(state \mid msg) = \frac{S_{\text{lit}}(msg \mid state)P(state)}{\sum_{state'} S_{\text{lit}}(msg \mid state')P(state')} \]

Literal speaker

\[ S_{\text{lit}}(msg \mid state) = \frac{\exp(\alpha (\log \lbrack msg, state \rbrack - C(msg)))}{\sum_{msg'} \exp(\alpha (\log \lbrack msg', state \rbrack - C(msg')))} \]
Pragmatic speakers

**Pragmatic speaker**

\[ S_{\text{prag}}(msg \mid state) = \text{pragmatic listener} - \text{message costs} \]

**Pragmatic listener**

\[ L_{\text{prag}}(state \mid msg) = \text{literal speaker} \times \text{state prior} \]

**Literal speaker**

\[ S_{\text{lit}}(msg \mid state) = \text{lexicon} - \text{message costs} \]
Joint inference

\[ L_{\text{prag}}(\text{state, Context} \mid \text{msg}) \]

\[ S_{\text{prag}}(\text{msg} \mid \text{state, Context}) \]
Limitations

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

\[ S_{\text{prag}}(msg \mid state) = \frac{\exp(\alpha (\log L_{\text{lit}}(state \mid msg) - C(msg)))}{\sum_{msg'} \exp(\alpha (\log L_{\text{lit}}(state \mid msg') - C(msg')))} \]

- High-bias model; few chances to learn from data
- Cognitive demands limit speaker rationality
- Speaker preferences
- Scalability
Neural RSA

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. A few other grounding ideas
Papers employing these techniques

- Andreas and Klein 2016
- Fried et al. 2018
- Monroe et al. 2017
- Monroe et al. 2018
Motivation

- Discriminative image labeling
- Image captioning
- Summarization
- Machine translation
- Collaborative problem solving
- Interpreting complex descriptions
- Optical Character Recognition

- Scalability
- Sensitivity to variation
- Bounded rationality
- New kinds of model assessment
- Impact
<table>
<thead>
<tr>
<th>Context</th>
<th>Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td>blue</td>
<td>blue</td>
</tr>
<tr>
<td>The darker blue one</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</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>

Stanford Colors in Context corpus (Monroe et al. 2017)
Literal neural speaker $S^\theta_{\text{lit}}$
Neural literal listener $\mathcal{L}_0$

Encoder

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

Decoder

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

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

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

- Fourier transform

- $f_1$, $f_2$, $f_T$

- $c_1$, $c_2$, $c_T$

softmax($s_1$, $s_2$, $s_T$)

Monroe et al. 2017
Neural pragmatic agents

Neural pragmatic speaker (Andreas and Klein 2016)

\[
S^\theta_{\text{prag}}(msg \mid state) = \frac{\mathcal{L}_0(state \mid msg)}{\sum_{msg' \in X} \mathcal{L}_0(state \mid msg')}
\]

with \(X\) a sample from \(S^\theta_{\text{lit}}(msg \mid state)\) such that \(msg \in X\).

Neural pragmatic listener

\[
\mathcal{L}_1(state \mid msg) \propto S^\theta_{\text{prag}}(msg \mid state)
\]

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.

Reasoning about all possible utterances/captions?
⇒ Sample from $S^\theta_{\text{lit}}$
⇒ 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 and Potts (2015): “RSA as a hidden activation function”

• Khani et al. (2018): Collaborative games with pragmatic reasoning.

• Cohn-Gordon and Goodman (2019): RSA for translation
Introspective speakers from Google

Generation and Comprehension of Unambiguous Object Descriptions
Junhua Mao2, Jonathan Huang1, Alexander Toshev1, Oana Camburu3, Alan Yuille2,4, Kevin Murphy1
1Google Inc., 2University of California, Los Angeles, 3University of Oxford, 4Johns Hopkins University
{mjhustc@,yuille@stat.}ucla.edu, oana-maria.camburu@cs.ox.ac.uk
{jonathanhuang,toshev,kpmurphy}@google.com

Context-aware Captions from Context-agnostic Supervision
Ramakrishna Vedantam1, Samy Bengio2, Kevin Murphy2, Devi Parikh3, Gal Chechik2
1Virginia Tech, 2Google, 3Georgia Institute of Technology
vrama91@vt.edu, parikh@gatech.edu, {bengio,kpmurphy,gal}@google.com

Mao et al. 2016; Vedantam et al. 2017
Google Refexp Dataset

Mao et al. 2016
Google Refexp Dataset examples

A boy brushing his hair while looking at his reflection.
A young male child in pajamas shaking around a hairbrush in the mirror.
The woman in black dress.
A lady in a black dress cuts a wedding cake with her new husband.
Zebra looking towards the camera.
A zebra third from the left.
A bird that is close to the baby in a pink shirt.
A bird standing on the shoulder of a person with its tail touching her face.

Mao et al. 2016
Maximum Mutual Information Training

Neural listener objective

Where an example is a message, a set of entities $I$, and a entity $ent \in I$:

$$J'(\theta) = - \sum_{n=1}^{N} \log \frac{S^\theta_{lit}(msg_n | ent_n; I_n)}{\sum_{ent' \in I_n} S^\theta_{lit}(msg_n | ent_n; I_n)}$$

Mao et al. 2016
Maximum Mutual Information Training

**Neural listener objective**

Where an example is a message, a set of entities $I$, and a entity $ent \in I$:

$$J'(\theta) = - \sum_{n=1}^{N} \log \frac{s_{\text{lit}}^\theta(msg_n | ent_n; l_n)}{\sum_{ent' \in l_n} s_{\text{lit}}^\theta(msg_n | ent_n; l_n)}$$

**Max margin objective**

To speed up training and make it more stable, they approximate the abovean max-margin objective that compares each target with a single randomly chosen distractor.

Mao et al. 2016
Introspective image captioners

**Target Image:**

![Image of a large passenger jet](image1.png)

**Distractor Image:**

![Image of a small airplane](image2.png)

**Speaker:**
An airplane is flying in the sky.

**Introspective Speaker:**
A large passenger jet flying through a blue sky.

Vedantam et al. 2017
Introspective speaker training

\[ \Delta(I, \text{state}, \text{state'}) = \arg\max_{msg} \left[ \lambda \log S_{\text{lit}}^\theta(msg | \text{state}; l_n) + (1 - \lambda) \log \frac{S_{\text{lit}}^\theta(msg | \text{state}; l_n)}{S_{\text{lit}}^\theta(msg | \text{state'}; l_n)} \right] \]

Proportional to a standard RSA \( \mathcal{L}_1 \).

Vedantam et al. 2017
Diagnosing the role of introspection

Target image and class
Rufous Hummingbird

Justifications vary with $\lambda$

- $\lambda = 0.00$: tarsals orange white brown wings wings orange tail dark an primaries
- $\lambda = 0.30$: This is a brown bird with a brown wing and a long pointy beak.
- $\lambda = 0.50$: This bird is brown with red on its neck and has a long, pointy beak.
- $\lambda = 0.70$: This is a bird with a white belly, brown wing and a red throat.
- $\lambda = 1.00$: A small sized bird that has a very long and pointed bill.

Distractor class
Ruby throated Hummingbird

Vedantam et al. 2017
Diagnosing the role of introspection

Target image and class
Rufous Hummingbird

Justifications vary with $\lambda$

- $\lambda = 0.00$ tarsals orange white brown wings wings orange tail dark an primaries
- $\lambda = 0.30$ This is a brown bird with a brown wing and a long pointy beak.
- $\lambda = 0.50$ This bird is brown with red on its neck and has a long, pointy beak.
- $\lambda = 0.70$ This is a bird with a white belly, brown wing and a red throat.
- $\lambda = 1.00$ context blind A small sized bird that has a very long and pointed bill.

Distractor class
Ruby throated Hummingbird

Vedantam et al. 2017
Diagnosing the role of introspection

Target image and class
Rufous Hummingbird

Justifications vary with $\lambda$

Distractor class
Ruby throated Hummingbird

- **fully discriminative**
  - $\lambda = 0.00$: tarsals orange white brown wings wings orange tail dark an primaries
  - $\lambda = 0.30$: This is a brown bird with a brown wing and a long pointy beak.
  - $\lambda = 0.50$: This bird is brown with red on its neck and has a long, pointy beak.
  - $\lambda = 0.70$: This is a bird with a white belly, brown wing and a red throat.
  - $\lambda = 1.00$: A small sized bird that has a very long and pointed bill.

Vedantam et al. 2017
Diagnosing the role of introspection

Target image and class
Rufous Hummingbird

Justifications vary with $\lambda$

- **fully discriminative**
  - $\lambda = 0.00$: tarsals orange white brown wings wings orange tail dark an primaries
  - $\lambda = 0.30$: This is a brown bird with a brown wing and a long pointy beak.
  - $\lambda = 0.50$: This bird is **brown with red on its neck** and has a long, pointy beak.
  - $\lambda = 0.70$: This is a bird with a **white belly**, **brown wing** and a **red throat**.
  - $\lambda = 1.00$: A small sized bird that has a very long and pointed bill.

Distractor class
Ruby throated Hummingbird

Vedantam et al. 2017
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
Grounded chat bots

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. A few other grounding ideas
Basic neural chatbot

Encoder

Decoder

not
much
</s>

not much
x1 x42 x39
h4 h5 h6

what’s
up
?

Decoder

not
much
</s>

x1 x42 x39
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 and 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 encoder GRU$_g$

Dialogue encoder GRU$_w$

Output encoder GRU$_o$

- $h^g$
- $h^s$
- Attention vector
- $h_1$, $h_2$, $h_3$, $h_4$
- $x_1$, $x_2$, $x_3$, $x_4$
- Item0=1, Item1=4, Item2=1
- $W_{item=0}$, $W_{item=1}$, $W_{item=2}$
Goal-based training

Goal encoder GRU<sub>g</sub>

Dialogue encoder GRU<sub>w</sub>

Output encoder GRU<sub>o</sub>

Agent A reads

Agent A writes

Agent B writes

Agent A reward
Decoding through rollouts

From 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/

- TRIPS
  http://www.cs.rochester.edu/research/cisd/projects/trips/

- TRAINS
  http://www.cs.rochester.edu/research/cisd/projects/trains/

- Cards
  http://CardsCorpus.christopherpotts.net/

- SCARE
  http://slate.cse.ohio-state.edu/quake-corpora/scare/

- The Carnegie Mellon Communicator Corpus
  http://www.speech.cs.cmu.edu/Communicator/
A few other grounding ideas

1. Overview: linguistic insights, and a bit of history
2. Speakers: From the world to language
3. Assignment/Bake-off overview: Speakers in context
4. Listeners: From language to the world
5. Reasoning about other minds
6. The Rational Speech Acts model (RSA)
7. Neural RSA
8. Grounded chat bots
9. A few other grounding ideas
Modeling users for sarcasm detection

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

Bla bla ... sentiment ... bla bla bla ... networks ...

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

I actually kind of liked it.

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

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”

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


References II


References


