Grounded Semantics

Berkeley

NL P

Daniel Fried
with slides from Greg Durrett and Chris Potts

Language is Contextual

- Some problems depend on grounding into perceptual or physical environments:
  - “Add the tomatoes and mix”
  - “Take me to the shop on the corner”

- The world only looks like a database some of the time!
- Most of today: these kinds of problems

Grounded Semantics

What things in the world does language refer to?

“Stop at the second car”

Pragmatics

How does context influence interpretation and action?

“Stop at the car”
Language is Contextual

• Some problems depend on grounding indexicals, or references to context
  • Deixis: “pointing or indicating”. Often demonstratives, pronouns, time
    and place adverbs
    • I am speaking
    • We won
    • He had rich taste
    • I am here
    • We are here
    • I’m in a class now
    • I’m in a graduate program now
    • I’m not here right now
  • (a team I’m on; a team I support)
  • (walking through the Taj Mahal)
  • (in my apartment; in this Zoom room)
  • (pointing to a map)
  • (note on an office door)

Some problems depend on grounding into speaker intents or goals:
  • “Can you pass me the salt”
    -> please pass me the salt
  • “Do you have any kombucha?” // “I have tea”
    -> I don’t have any kombucha
  • “The movie had a plot, and the actors spoke audibly”
    -> the movie wasn’t very good
  • “You’re fired!”
    -> performative, that changes the state of the world
  • More on these in a future pragmatics lecture!

Language is Contextual

• Some knowledge seems easier to get with grounding:

  Winograd schemas
  The large ball crashed right through the table because it
  was made of steel. What was made of steel?
  -> ball

  The large ball crashed right through the table because it
  was made of styrofoam. What was made of styrofoam?
  -> table

  “blinking and breathing problem”

<table>
<thead>
<tr>
<th>Word</th>
<th>Temporary</th>
<th>Exact</th>
<th>Word</th>
<th>Temporary</th>
<th>Exact</th>
</tr>
</thead>
<tbody>
<tr>
<td>sprit</td>
<td>15,775,917</td>
<td>86,492</td>
<td>biggest</td>
<td>110,000</td>
<td>15,576</td>
</tr>
<tr>
<td>bangl</td>
<td>3,968,793</td>
<td>165,347</td>
<td>biggest</td>
<td>3,966</td>
<td>3,824</td>
</tr>
<tr>
<td>adorned</td>
<td>2,885,529</td>
<td>11,208</td>
<td>tan-law</td>
<td>368,222</td>
<td>10,308</td>
</tr>
<tr>
<td>school</td>
<td>988,693</td>
<td>4,832</td>
<td>excluded</td>
<td>648,985</td>
<td>5,981</td>
</tr>
<tr>
<td>armed</td>
<td>723,036</td>
<td>34,912</td>
<td>tan-bridged</td>
<td>5,887</td>
<td>311</td>
</tr>
</tbody>
</table>

  Table 1: Frequencies from [3] and the number of times Kesten
  learns that A person may not be a person, including appropriate
  arguments, e.g., A person may be a person. For instance, more
  frequently encountered in the passive, we include be murdered.

  Winograd 1972; Levesque 2013; Wang et al. 2018
  Gordon and Van Durme, 2013


  Children learn word meanings incredibly fast, from incredibly few data
  • Regularity and contrast in the input signal
  • Social cues
  • Inferring speaker intent
  • Regularities in the physical environment
Grounding

- (Some) possible things to ground into:
  - **Percepts**: red means this set of RGB values, loud means lots of decibels on our microphone, soft means these properties on our haptic sensor...
  - **High-level precepts**: cat means this type of pattern
  - **Effects on the world**: go left means the robot turns left, speed up means increasing actuation
  - **Effects on others**: polite language is correlated with longer forum discussions

Grounding

- (Some) key problems:
  - **Representation**: matching low-level percepts to high-level language (pixels vs cat)
  - **Alignment**: aligning parts of language and parts of the world
  - **Content Selection / Context**: what are the important parts of the environment to describe (for a generation system) or focus on (for interpretation)?
  - **Balance**: it’s easy for multi-modal models to “cheat”, rely on imperfect heuristics, or ignore important parts of the input
  - **Generalization**: to novel world contexts / combinations

Grounding

- Today, survey:
  - Spatial relations
  - Image captioning
  - Visual question answering
  - Instruction following

Spatial Relations
Spatial Relations

‣ How would you indicate O1 to someone with relation to the other two objects? (not calling it a vase, or describing its inherent properties)
‣ What about O2?
‣ Requires modeling listener — “right of O2” is insufficient though true

Spatial Relations

‣ Two models: a speaker, and a listener
‣ We can compute expected success:

\[ EU(S, L) = \sum_{o,w,g} p(o)p_s(w|o)p_s(g|w)U(o, g) \]

‣ Modeled after cooperative principle of Grice (1975): listeners should assume speakers are cooperative, and vice-versa
‣ For a fixed listener, we can solve for the optimal speaker, and vice-versa

Spatial Relations

‣ Listener model:
‣ Objects are associated with coordinates (bounding boxes of their projections). Features map lexical items to distributions (“right” modifies the distribution over objects to focus on those with higher x coordinate)
‣ Language -> spatial relations -> distribution over what object is intended

Spatial Relations

‣ Listener model:
‣ Syntactic analysis of the particular expression gives structure
‣ Rules (O2 = 100% prob of O2), features on words modify distributions as you go up the tree
Spatial Relations

‣ Put it all together: speaker will learn to say things that evoke the right interpretation
‣ Language is grounded in what the speaker understands about it

Image Captioning

How do we caption these images?

‣ Need to know what’s going on in the images — objects, activities, etc.
‣ Choose what to talk about
‣ Generate fluid language

Pre-Neural Captioning: Objects and Relations


‣ Detect objects using (non-neural) object detectors trained on a separate dataset
‣ Label objects, attributes, and relations. CRF with potentials from features on the object and attribute detections, spatial relations, and text co-occurrence
‣ Convert labels to sentences using templates
**ImageNet models**

- ImageNet dataset (Deng et al. 2009, Russakovsky et al. 2015)
  - Object classification: single class for the image. 1.2M images, 1000 categories
  - Object detection: bounding boxes and classes. 500K images, 200 categories
- 2012 ImageNet classification competition: drastic error reduction from deep CNNs
- Last layer is just a linear transformation away from object detection — should capture high-level semantics of the image, especially what objects are in there

**Neural Captioning: Encoder-Decoder**

- Use a CNN encoder pre-trained for object classification (usually on ImageNet). Freeze the parameters.
- Generate captions using an LSTM conditioning on the CNN representation

**What’s the grounding here?**

- a close up of a plate of ___
- a couple of bears walking across ___
- What are the vectors really capturing?
  - Objects, but maybe not deep relationships

**Simple Baselines**

- MRNN: take the last layer of the ImageNet-trained CNN, feed into RNN
- k-NN: use last layer of the CNN, find most similar train images based on cosine similarity with that vector. Obtain a consensus caption.

<table>
<thead>
<tr>
<th>LM</th>
<th>PPLX</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA-NET</td>
<td>11.3</td>
<td>22.0</td>
<td>22.6</td>
</tr>
<tr>
<td>DA-ST</td>
<td>14.3</td>
<td>22.4</td>
<td>22.6</td>
</tr>
<tr>
<td>MRNN</td>
<td>13.2</td>
<td>25.7</td>
<td>22.6</td>
</tr>
</tbody>
</table>

Table 1: Model performance on testvil. ⊥: From (Fang et al., 2015).

Devlin et al. (2015)
**Simple Baselines**

<table>
<thead>
<tr>
<th>System</th>
<th>Unique Captions</th>
<th>Seen In Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>99.4%</td>
<td>4.8%</td>
</tr>
<tr>
<td>D-ME+DMSM</td>
<td>33.4%</td>
<td>40.4%</td>
</tr>
<tr>
<td>MRNN</td>
<td>33.1%</td>
<td>60.3%</td>
</tr>
<tr>
<td>D-ME+DMSM+MRNN</td>
<td>28.5%</td>
<td>61.3%</td>
</tr>
<tr>
<td>k-Nearest Neighbor</td>
<td>36.6%</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Table 6: Percentage unique (Unique Captions) and novel (Seen In Training) captions for testset images. For example, 28.5% unique means 5,776 unique strings were generated for all 20,244 images.*

- Even from CNN+RNN methods (MRNN), relatively few unique captions even though it’s not quite regurgitating the training

---

**Neural Captioning: Object Detections**

- Follow the pre-neural object-based systems: use features predictive of individual objects and their attributes

*Training data (Visual Genome, Krishna et al. 2015):*

*Object and attribute detections (Faster R-CNN, Ren et al. 2015):*

---

**Neural Captioning: Object Detections**

- Also add an attention mechanism: attend over the visual features from individual detected objects

---

**Neural Hallucination**

- Language model often overrides the visual context:

  - A group of people sitting around a *table* with laptops
  - A kitchen with a stove and a *sink*

- Standard text overlap metrics (BLEU, METEOR) aren’t sensitive to this!
Visual Question Answering

- Answer questions about images
- Frequently require compositional understanding of multiple objects or activities in the image

Visual Question Answering

- What is in the sheep's ear? => tag

Neural Module Networks

- Integrate compositional reasoning + image recognition
- Have neural network components like `find [sheep]` whose composition is governed by a parse of the question
- Like a semantic parser, with a learned execution function

Neural Module Networks

- Fuse modalities: pre-trained CNN processing of the image, RNN processing of the language
- What could go wrong here?

Agrawal et al. (2015)

Agrawal et al. (2015)

Human-written questions

CLEVR: Johnson et al. (2017)

Synthetic, but allows careful control of complexity and generalization

Agrawal et al. (2015)

Visual Module Networks

What size is the cylinder that is left of the brown metal thing that is left of the big sphere?

VQA: Agrawal et al. (2015)

Andreas et al. (2016), Hu et al. (2017)
Neural Module Networks

- Able to handle complex compositional reasoning, at least with simple visual inputs

Andreas et al. (2016), Hu et al. (2017)

Visual Question Answering

- In many cases, language as a prior is pretty good!
  - “Do you see a…” = yes (87% of the time)
  - “How many…” = 2 (39%)
  - “What sport…” = tennis (41%)
- When only the question is available, baseline models are super-human!
- Balanced VQA: reduce these regularities by having pairs of images with different answers

Goyal et al. (2017)

Challenge Datasets

- NLVR2: Difficult comparative reasoning; balanced dataset construction; human-written

<table>
<thead>
<tr>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>One image contains a single rabbit in a standing pose with its head and body facing leftward, and the other image contains a group of at least eight rabbits.</td>
<td>There are two strings in total traveling in the same direction.</td>
</tr>
</tbody>
</table>

Table 3: Six examples with three different sentences from NLVR2. For each sentence, we show two examples using different image pairs, each with a different label.

Suhr & Zhou et al., 2019

Instruction Following

Majority class baseline: 50%
Current best system: 80%
Human performance: 96%
Instruction Following

- SAIL dataset: navigational instructions in synthetic grid worlds, with furniture and patterns
  MacMahon et al., 2006; Chen and Mooney, 2011

Human annotator view

System view

Instruction Following

Input instruction: go to the chair. turn left and go forward to the fish painting. head to the right until you get to a coat rack

Output actions:

Instruction Following

- Several successful approaches using semantic parsing
  (Chen and Mooney 2011; Artzi and Zettlemoyer 2013; Artzi et al. 2014)

Input instruction: go to the chair. turn left and go forward to the fish painting. head to the right until you get to a coat rack

Output actions:
Several successful approaches using semantic parsing (Chen and Mooney 2011; Artzi and Zettlemoyer 2013; Artzi et al. 2014)

- Logical forms denote action sequences, often using post-conditions
- Learn from action sequences paired with instructions

Examples from Yoav Artzi

```
go to the chair
\( \lambda a.\text{move}(a) \land \text{to}(a, \text{sz.chair}(x)) \)
```

```
move until you reach the chair
\( \lambda a.\text{move}(a) \land \\
\text{post}(a, \text{intersect(sz.chair}(x), \text{you})) \)
```

Neural Instruction Following

- Encoder-decoder setup with attention to the instruction
- Decoder takes as input embeddings for all the (symbolic) world features the agent can see

```
LSTM encoder
```

```
LSTM decoder with attention
```

- Almost as good as the best semantic parsing approach

Mei et al. (2016)

Vision-and-Language Navigation

```
Turn left and take a right at the table. Take a left at the painting and then take your first right. Wait next to the exercise equipment.
```

Anderson et al. (2018)
Discrete motion, but real images

Vision-and-Language Navigation

Anderson et al. (2018)
Vision-and-Language Navigation

- LSTM Encoder
- LSTM Decoder with Attention

Anderson et al. (2018)

Walk past hall table. Walk into bedroom. Make left at table clock. Wait at bathroom door threshold.

Fried, Hu, Cirik et al. (2018)

- Best current models: 72% accuracy; humans: 86%
- But, what are the models actually grounding into?
- Some combination of:
  - generalizable representations
  - environments seen in training
  - biases in the routes themselves

Gordon et al. 2018, Hu et al. 2019
Challenge Tasks

Touchdown
Chen et al. 2019, Mehta et al. 2020

- Long, complex routes through NYC’s StreetView graph, with associated imagery
- SOTA model: 5% accuracy. Human: 92%

ALFRED
Shridhar et al. 2020

- Interact with objects in a household setting
- Long time horizons, non-reversible state changes
- Baseline model: 1% accuracy. Human: 91%

Takeaways

- Lots of problems where natural language has to be interpreted in an environment and can be understood in the context of that environment
- Neural models make it easier to fuse representations from multiple modalities (but they sometimes learn to cheat)
- Symbolic methods guided by linguistic structure; neural systems with learned representations; some work productively combines both