Grounded Semantics

Berkeley

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”
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* (a team I’m on; a team I support)  
  - *He had rich taste* (walking through the Taj Mahal)  
  - *I am here* (in my apartment; in this Zoom room)  
  - *We are here* (pointing to a map)  
  - *I’m in a class now*  
  - *I’m in a graduate program now*  
  - *I’m not here right now* (note on an office door)
Language is Contextual

- 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>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>spoke</td>
<td>11,577,917</td>
<td>244,458</td>
</tr>
<tr>
<td>laughed</td>
<td>3,904,519</td>
<td>169,347</td>
</tr>
<tr>
<td>murdered</td>
<td>2,843,529</td>
<td>11,284</td>
</tr>
<tr>
<td>inhaled</td>
<td>984,613</td>
<td>4,412</td>
</tr>
<tr>
<td>breathed</td>
<td>725,034</td>
<td>34,912</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word</th>
<th>Teraword</th>
<th>Knext</th>
</tr>
</thead>
<tbody>
<tr>
<td>hugged</td>
<td>610,040</td>
<td>10,378</td>
</tr>
<tr>
<td>blinked</td>
<td>390,692</td>
<td>20,624</td>
</tr>
<tr>
<td>was late</td>
<td>368,922</td>
<td>31,168</td>
</tr>
<tr>
<td>exhaled</td>
<td>168,985</td>
<td>3,490</td>
</tr>
<tr>
<td>was punctual</td>
<td>5,045</td>
<td>511</td>
</tr>
</tbody>
</table>

Table 1: Frequencies from [3] and the number of times Knext learns that *A person may* *(x)*, including appropriate arguments, e.g., *A person may hug* *a person.* For murder, more frequently encountered in the passive, we include *be murdered.*

Winograd 1972; Levesque 2013; Wang et al. 2018

Gordon and Van Durme, 2013
Language is Contextual

- 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

Golland et al. (2010)
Spatial Relations

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

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

- 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 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?

- 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>D-ME†</td>
<td>18.1</td>
<td>23.6</td>
<td>22.8</td>
</tr>
<tr>
<td>D-LSTM</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>
<tr>
<td>k-Nearest Neighbor</td>
<td>-</td>
<td>26.0</td>
<td>22.5</td>
</tr>
<tr>
<td>1-Nearest Neighbor</td>
<td>-</td>
<td>11.2</td>
<td>17.3</td>
</tr>
</tbody>
</table>

Table 1: Model performance on testval. †: 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>47.0%</td>
<td>30.0%</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 testval 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

Devlin et al. (2015)
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):

Anderson et al. (2018)
Neural Captioning: Object Detections

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

Anderson et al. (2018)
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!

Slide credit: Anja Rohrbach

Rohrbach & Hendricks et al. (2018)
Visual Question Answering
Visual Question Answering

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

VQA: Agrawal et al. (2015)
Human-written questions

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

CLEVR: Johnson et al. (2017)
Synthetic, but allows careful control of complexity and generalization
Visual Question Answering

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

Agrawal et al. (2015)
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

What is in the sheep’s ear? => tag

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)
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
**Challenge Datasets**

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

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

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

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

\[
\begin{align*}
S & \quad \text{go} \\
AP/NP & \quad \lambda a. \text{move}(a) \\
NP/N & \quad \lambda x. \lambda a. \text{to}(a, x) \\
N & \quad \lambda f. \lambda x. \text{f}(x) \\
N & \quad \lambda x. \text{chair}(x) \\
NP & \quad \text{the} \quad \lambda x. \text{chair}(x) \\
ix & \quad \text{chair} \\
NP & \quad \text{to} \quad \lambda x. \text{chair}(x) \\
AP & \quad \lambda a. \text{to}(a, \text{ix.chair}(x)) \\
S\backslash S & \quad \lambda f. \lambda a. \text{f}(a) \land \text{to}(a, \text{ix.chair}(x)) \\
S & \quad \text{go} \land \text{to}(a, \text{ix.chair}(x))
\end{align*}
\]

examples from Yoav Artzi
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
Instruction Following

- This is a sequence-to-sequence task, right?

Inputs

- go forward to the grey hallway

Instruction

Listener

Outputs

Actions in context
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

- Almost as good as the best semantic parsing approach

Mei et al. (2016)
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

LSTM Encoder

LSTM Decoder with Attention

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

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

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

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

Anderson et al. (2018)
Walk past hall table. Walk into bedroom. Make left at table clock. Wait at bathroom door threshold.
Vision-and-Language Navigation

- 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%

*Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic lights...*
Challenge Tasks

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

ALFRED Shridhar et al. 2020
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