Recall: Seq2seq Chatbots

Just like conventional MT, can train seq2seq models for this task
Why might this model perform poorly? What might it be bad at?

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td>0.33</td>
</tr>
<tr>
<td>MT</td>
<td>3.21</td>
</tr>
<tr>
<td>HUMAN</td>
<td>6.08</td>
</tr>
</tbody>
</table>

Recall: Lack of Diversity

Solution: mutual information criterion; response R should be predictive of user utterance U as well

Standard conditional likelihood: \( \log P(R|U) \)

Mutual information: \( \log \frac{P(R,U)}{P(R)P(U)} = \log P(R|U) - \log P(R) \)

\( \log P(R) \) can reflect probabilities under a language model

Li et al. (2016)
Recall: Task-Oriented Dialogue

- Task 1: Issuing API calls
  - Hi! Hello, what can I help you with today?
  - I'd like to book a table for six people in an expensive price range with British food.
  - I'm on it! Where should it be?
  - In London.
  - Ok, let me look into some options for you.
  - Actually I would prefer for four.
  - Sure. Is there anything else to update?
  - No.
  - Ok let me look into some options for you.

- Task 2: Updating API calls
  - api_call(British, London, Six, Expensive)
  - api_call(British, London, Four, Expensive)

Using either wizard-of-Oz or other annotations, can collect static traces and train from these

Bordes et al. (2017)

Recall: QA as Dialogue

- Dialogue is a very natural way to find information from a search engine or a QA system

- Several recent datasets on this topic, but tough to collect a static dataset for an interactive application

Iyyer et al. (2017)

This Lecture

- Example grounding applications
- Image captioning / VQA
- Grounding with interaction

Basic Grounding Examples
History

- Miller and Johnson-Laird (1976) — Language and Perception
- Harnad (1990) — Symbol grounding problem
  - How do we connect “symbols” to the world in the right way?

In a pure symbolic model the crucial connection between the symbols and their referents is missing; an autonomous symbol system, though amenable to a systematic semantic interpretation, is ungrounded. In a pure connectionist model, names are connected to objects through invariant patterns in their sensory projections, learned through exposure and feedback, but the crucial compositional property is missing; a network of names, though grounded, is not yet amenable to a full systematic semantic interpretation. In the hybrid system proposed here, there is no longer any autonomous symbolic level at all; instead, there is an intrinsically dedicated symbol system, its elementary symbols (names) connected to nonsymbolic representations that can pick out the objects to which they refer, via connectionist networks that extract the invariant features of their analog sensory projections.
- Neural networks (connectionism) help us connect symbolic reasoning to sensory inputs

Grounding

- Tie language to something concrete in the world
  - Percepts: red means this set of RGB values, loud means lots of decibels on our microphone, soft means these properties on our haptic sensor...
  - Higher-level percepts: cat means this type of pattern in an image
  - Effects on others: go left means the robot turns left, speed up means increasing actuation

Colors

- What color is this?
- What about this?

McMahan and Stone (2014)

Colors

- When we say “yellowish-green”, what does that mean?
- Color descriptions governed by perception as well as availability: how commonly it is used (yellowish green vs. chartreuse)

McMahan and Stone (2014)
**Colors**

- \( P(k_{\text{true}} \mid X) \): distribution parameterized in HSV space as follows: there are certain ranges where a color can “definitely apply”, others where it can apply
- \( P(k_{\text{said}} \mid k_{\text{true}}) \): captures availability; prior towards common colors
- Model combines language / reasoning with basic perception — characteristic of grounding

Mcmahan and Stone (2014)

**Spatial Relations**

- Grice (1975)
  - Maxim of quality: say something true
  - Maxim of quantity: be as informative as required but no more
  - Maxim of relation: be relevant
  - Maxim of manner: avoid ambiguity
  - Maximize expected utility given listener model
    \[
    EU(s, l) = \sum_{o,w,g} p(o)p_s(w \mid o)p_l(g \mid w)U(o, g)
    \]
  - Say something which has a high probability of evoking the right response in the listener

Golland et al. (2010)

- 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

Golland et al. (2010)
Objects are associated with coordinates, 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

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

Want to be able to follow instructions in a virtual environment

“Go along the blue hall, then turn left away from the fish painting and walk to the end of the hallway”

Basic plans derived directly from supervision

“Landmarks” plans — things that should be true after each step (which may show up in the language)
Instruction Following

- Train semantic parser on (utterance, action) pairs
- Language is grounded in actions in the world

Chen and Mooney (2011)

Instruction Following

- “Spatial description clauses” -> “grounding graphs”
- Commands from the corpus:
  - Go to the first crate on the left and pick it up.
  - Pick up the pallet of boxes in the middle and place them on the trailer to the left.
  - Go forward and drop the pallets to the right of the first set of tires.
  - Pick up the tire pallet off the truck and set it down

Tellex et al. (2011)

Connections to Semantic Parsing

- Each grounding framework requires mapping natural language to something concrete (distribution in color space, object, action sequence)
- Sometimes looks like semantic parsing, particularly when language -> discrete output
- Using linguistic structure to capture compositionality is often useful

Image Captioning
How do we caption these images?

- Need to know what’s going on in the images — objects, activities, etc.

ImageNet models

- Train on ImageNet to do object classification

  - 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

ImageNet models

- Many architectures for this: VGG, ResNet, DenseNet, etc. — all end in fully-connected layers

Images -> Text

- English Sentence $\rightarrow$ CNN Encoder $\rightarrow$ RNN encoder $\rightarrow$ RNN decoder $\rightarrow$ French Sentence [Sutskever et al. NIPS'14]
- [Donahue et al. CVPR'15]
- [Venugopal et al. ICCV'15]
- [Vinyals et al. CVPR'15]
What’s the grounding here?

- a close up of a plate of ___
- a couple of bears walking across ____

What are the vectors really capturing? Probably some objects, but maybe not deep relationships

Simple Baselines

- Simple baselines work well!
- **D-**: condition on detections only
- **MRNN**: take the last layer of the CNN, feed into RNN
- **k-NN**: use last layer of ImageNet model, find most similar train images based on cosine similarity with that vector

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 3,776 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)

Video Captioning

- Generate an NL video description by training a suite of SVM-based visual recognizers and composing their outputs into a coherent sentence using a graphical model

(Krishnamoorthy et al., 2013; Thomason et al., 2014)

Video

- Content Selection
- SVOP
- Surface Realization
- Sentence

Slide credit: Ray Mooney
Video Captioning

MAP inference on Factor Graph estimates the most likely SVOP quadruple.

Subject: person
Verb: slice
Object: onion
Place: kitchen

A person is slicing an onion in the kitchen.

Visual Question Answering

- Answer questions about images
- Frequently much more metaphorical, require compositional understanding of multiple objects + activities in the image

Agrawal et al. (2015)

Visual Question Answering

- CNN processing of the image, RNN processing of the language
- What could go wrong here?

Neural Module Networks

- Integrate compositional reasoning + image recognition
- Have neural network components like classify[color] whose use is governed by a parse of the question

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

- Can also learn these structures with reinforcement learning

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%)
- Balanced VQA: remove these regularities by having pairs of images with different answers

Goyal et al. (2017)

Understanding VQA


Q: What is the person doing?  A: Skiing

Explanation: “Because he is on a snowy hill wearing skis”

Grounding Language in Interaction
Grounding in Interaction

- Corpus of dialogues — can train a model on these to learn to negotiate

Grounding in Interaction

- Same issues as other dialogue systems: system may prefer generic choices, like accepting the offer, instead of negotiating harder
- Instead: do self-play rollouts, train with reinforcement learning to maximize reward and not likelihood of human utterances

Grounding in Interaction

- Interleave self-play with supervised learning, otherwise the messages stop looking like real English
- When two systems talk to each other, they remap what words mean and completely change the grounding

Grounding in Interaction

- Less direct form of grounding: we understand the language used based on the effects it produces in the other agent (whether human or machine) and in the final reward
- More “symbolic” than grounding percepts like color, but still about interacting with the world!
Takeaways

- Lots of problems where natural language has to be interpreted in an environment and can be understood in the context of that environment.

- Image recognition: particularly large area of research featuring big neural networks (but they sometimes learn to cheat).

- More complex environments/robots/simulations/tasks -> more complex dialogue to be learned over time!