Learning and using language via recursive pragmatic reasoning about other agents

Our domain: We simulate language usage in simple referential games

Our problem: What is a word's meaning?

In existing learning models:

- a communicative goal, requiring recursive reasoning

- a latent variable

- Word
  - Lexicon
  - Referent

In existing pragmatics models:

- Acts like $L_0$, a communicative goal requiring recursive reasoning
- Given the speaker produced this word, what do they desire me to believe?
- $P_{S_0}(w|o) = \frac{P_{S_0}(o|w)P(w)}{P(o)}$  

- Acts like $L_1$, a communicative goal requiring recursive reasoning
- What word will produce the desired belief in my listener?
- $P_{S_1}(o|w) = \frac{P_{S_1}(w|o)P(o)}{P(w)}$

A fundamental paradox for social learning

Word meaning, in the latent variable sense, does not appear in either Bayesian pragmatics models or the real world's actual generative process. Without uncertainty about a latent variable, learning is impossible. Yet humans both learn word meanings and perform pragmatic reasoning. How can we reconcile these approaches?

Several obvious solutions don't work

- Make $L_0$ or $S_0$ uncertain about the literal lexicon?
  - But then actual speaker and listener must both marginalize out this uncertainty, so no-one's behavior is sensitive to the actual literal lexicon, so there is no data to learn it.
  - Make each recursive agent maintain uncertainty about their model of the next agent down?
  - Arguably correct in theory, but would require actual speaker and listener to learn hyper-hyper-…-hyper-distributions, which is probably impossible even in principle due to data sparsity.

Our solution: Assume conventionality + knowledgeable peers

Each agent assumes:
(a) There is a specific, "conventional" literal lexicon that everyone is supposed to be using, but everyone else knows what this lexicon is, and believes that I know it as well,
(b) but in fact I don't know it, and have to do my best to fake it.
Assumption (a) explains why naive language users will argue – falsely! – that words have objective meanings (the "lexicographer's illusion"). Assumption (b) means that data is available, but avoids the explosion of hyper-distributions (and is uncomfortably familiar). Assumption (c) means there's something to learn. When combined with standard Bayesian techniques, these assumptions give a definition of what a "convention" is, and provide a mechanism for learning, using, and creating them.

Our model

- Convention-based speaker $S$: Acts like $S_0 +$ uncertainty about $L_0$.
  - $P_{S}(w|o) = \frac{P_{S}(o|w)P(w)}{P(o)}$  
  - Learning: uses $L_1$ as generative model

- Convention-based listener $L$: Acts like $L_1 +$ uncertainty about $L_0$.
  - $P_{L}(w|o) = \frac{P_{L}(o|w)P(w)}{P(o)}$  
  - Learning: uses $S_0$ as generative model

Learning scalar quantifiers like "some" and "all"

Training data: pragmatically strengthened uses

<table>
<thead>
<tr>
<th>Marginal literal belief</th>
</tr>
</thead>
<tbody>
<tr>
<td>Districhlet-multinomial</td>
</tr>
</tbody>
</table>

Emergence of novel & efficient conventions in interaction

When humans interact, novel and task-adapted communicative systems emerge. We simulate interactions between agents who begin with uniform priors over lexicons.

<table>
<thead>
<tr>
<th>Mean communicative success over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dialogue turn</td>
</tr>
</tbody>
</table>

Example runs

- Better run: Dialogue turn
  - "some*" 0.78 0.22
  - "all*" 0.78 0.22

- Bias towards "Horn-compatible" lexicons; implicature shifts to become literal meaning

- Biases from "Irrational" social anxiety assumption systematically drive the communicative system towards greater efficiency, and in the long run may leave their mark on structure of languages themselves.

References: