Language understanding and Bayesian inference

**Topic 4: Bayesian pragmatics**

Daniel Lassiter

Stanford Linguistics

NASSLLI ‘14

UMD
hat tips

- Mike Frank
- Adam Vogel
- Chris Potts
course plan

1. probability and Bayesian inference
2. noisy channels, audience design, modularity & Bayesian inference
3. anchoring interpretation in world models
4. Bayesian pragmatics
today

- architecture for recursive pragmatics
- quantity implicature
- context-dependence and gradable adjectives
Pragmatic enrichment

“I recommend this candidate for your position. He has great penmanship.”

=> He’s not qualified, but the author is too polite to say so outright

“Some of the students passed the test”

=> Not all did
Gricean pragmatics

Speakers assumed to obey maxims
  – Quality: be truthful
  – Quantity: be informative
  – Relation: be relevant
  – Manner: be concise, clear

Listeners assume speakers follow maxims
  – E.g. couldn’t say that all students passed
  – E.g. couldn’t say more (positive) things about recommendee [??]

“One of my avowed aims is to see talking as a special case or variety of purposive, indeed rational, behavior…”

Grice (1975)
pragmatics as action understanding

“Grice’s maxims taken collectively mean ‘Don’t include elements that don’t do anything.’ Our position is that, under a goal-oriented view of language... there is no need to explicitly follow such a directive at all; the desired behavior just falls out of the mechanism.”

• Horn (1984)
  – Q Principle: Make your contribution sufficient
  – R Principle: Say no more than you must

• balance informativity and cost

• speech as goal-driven (‘rational’) behavior

• interpretation depends on this assumption

Dale & Reiter (1996)
recursive pragmatics

choose a starting point, proceed by simulation

1. Literal interpretation uses a learned grammar and lexicon

2. Ss choose utterances by simulating literal listener

3. Pragmatic Ls infer $b_S, d_S$ by simulating Ss

    [...] further iterations? ...]

(Goodman & Lassiter ’14, cf. Lewis ’69, Franke ’09, ’12, etc.)
def literal-listener(utt, model, QUD):
    return query(model, QUD, meaning(utt) == True)

def speaker(answer, model, QUD):
    return query(model, utt, utt == utt-prior() and
                  answer(model, QUD) == lit-listener(utt, listener-model, QUD))

def pragmatic-listener(utt, model, QUD):
    val = answer(model, QUD)
    return query(model, QUD,
                 utt == speaker(val, speaker-model, QUD))
recursive pragmatics

\[ P_{L_0}(A|u) = P_{L_0}(A|[u] = 1) \]

\[ P_{S_1}(u|A) \propto P_{L_0}(A|u) \times P_{S_1}(u) \]

alternatives come in here

\[ P_{L_1}(A|u) \propto P_{S_1}(u|A) \times P_{L_1}(A) \]
decision-theoretic formulation

\[ \mathbb{U}_S(u; A) \propto \log_2(P_{L_0}(A|u)) - C(u) \]

Luce choice rule:

\[ P_{S_1}(u|A) \propto \exp(\alpha \times \mathbb{U}(u; A)) \]

models equivalent with alpha = 1
quantity implicature

• What’s the weather like? It’s warm.
• How many students passed? Some did.
• Anything interesting on your trip? I met a girl.
• How’s dinner prep going? I’ve washed the lettuce.

My friend has glasses.
SI: literal listener

“Jane played in some match”.

• Suppose there were 3 matches total.

• How many did she play in?
“Jane played in some match”.

SI: speaker

![Graph showing probability of choosing different utterances based on number of matches Jane played in, with lines for "None", "Some", and "All" choices.](image-url)
SI: speaker

Pragmatic listener

Literal listener

Speaker

Prior

Some

All

Probability that speaker chooses utterance

Probability according to literal listener

Probability according to reflective listener

Number of matches Jane played in

Number of matches Jane played in
ad hoc SI

Quantifier Scale

“some” ( )

“all” ( )

Ad-hoc Scale (Experimental Stimulus)

(Stiller et al. 2014)
alternatives

• alternatives matter in $P_S(u|A)$ normalization
• utterance prior relevant: e.g. length
  – some $=\not=\Rightarrow$ not(some but not all)
• probably not all possibilities considered
  – computational or processing explanation?
DialogBot

Spatial language, CARDS task (Potts 2012)
  – http://cardscorpus.christopherpotts.net/
reinforcement learning: multi-agent POMDPs
  – approximate belief representations

(Vogel et al., 2013; cf. Golland et al. 2010)
Exact Multi-agent Belief Update
Approximate Multi-agent Belief Update
DialogBot

- Quality, relevance emerge due to shared reward function (still to test: manner, quantity …)
- behavior gets more Gricean with experience
Relative adjectives

- heavy, tall, happy, expensive, ...
- context-dependent meaning
  - tall \{boy, NBA player, tree, building\}
  - expensive \{coffee, shirt, house, airplane\}
  - heavy \{child, rock, car, planet\}
- interpretation depends on distribution of a property in some reference class
Relative adjectives

- **heavy, tall, happy, expensive, ...**
- no sharp boundaries
- borderline cases
- **sorites paradox:**
  - A house that costs $10,000,000 is expensive.
  - A house $1 cheaper than an expensive house is also expensive.
  - Therefore, a house that costs $1 is expensive.
Absolute adjectives

• Puzzlingly different: full/empty, wet/dry, safe/dangerous, ...
  • meanings are less (not?) context-dependent
  • meanings are sharp(er)
  • reference classes apparently not relevant to interpretation
  • sorites paradox not compelling
Semantics of GAs

schema

\[[A] = \lambda \theta_A \lambda x [\mu_A(x) > \theta_A]\n
examples

\[[\text{expensive}] = \lambda \theta_{\text{exp}} \lambda x_e [\mu_{\text{exp}}(x) > \theta_{\text{exp}}]\n
\[[\text{tall}] = \lambda \theta_{\text{tall}} \lambda x [\mu_{\text{height}}(x) > \theta_{\text{tall}}]\n
\[[\text{full}] = \lambda \theta_{\text{full}} \lambda x [\mu_{\text{fullness}}(x) > \theta_{\text{full}}]\n
variable-free

\[[\text{Al is tall}] = \lambda \theta_{\text{tall}} [\mu_{\text{tall}}(A) > \theta_{\text{tall}}]\n
POS-based

\[[\text{Al is POS tall}]^{\theta_{\text{tall}}} = \mu_{\text{tall}}(A) > \theta_{\text{tall}}]
In Church

\[
\text{lexicon('expensive')} \Rightarrow \lambda d \lambda x . \text{cost}(x) > d
\]
\[
\text{lexicon('strong')} \Rightarrow \lambda d \lambda x . \text{strength}(x) > d
\]
\[
\text{lexicon('POS')} = \lambda A \lambda x . A(x) > ..... ??
\]
\[
\text{meaning('POS strong')} \Rightarrow \lambda x . \text{strength}(x) > ..... ??
\]

- The threshold value comes from context.
- What is ‘context’?
- Bayesian perspective: not given, but inferred.
def literal-listener(utt, model, QUD):
    index = index-prior()
    return query(model, QUD, meaning(utt, index) = True)

u = “Al is strong”, QUD= “How strong?”
What will this model infer?

def prag-listener(utt, model, QUD):
    val = answer(model, QUD)
    return query(model, QUD, utt == speaker(val, speaker-model, QUD)

def speaker(answer, model, QUD):
    return query(model, utt, utt=utt-prior() and answer(model, QUD) ==
    lit-listener(utt, listener-model, QUD)
Disaster

“Al is strong” conveys no information about Al’s height.

literal-listener is only under pressure to make utterance true. Result is preference for very weak meanings.
def literal-listener(utt, model, QUD, index):
    return query(model, QUD, meaning(utt, index) = True)

def speaker(answer, model, QUD, index):
    return query(model, QUD, answer(model, QUD) ==
                 lit-listener(utt, listener-model, QUD, index))

def prag-listener(utt, model, QUD):
    index = index-prior()
    val = answer(model, QUD)
    return query(model, QUD, utt ==
                 speaker(val, speaker-model, QUD, index))

Lifted indices

• index fills in “strong” threshold $\theta$

• speaker’s goal is to get listener to choose correct answer to QUD

• speaker is more likely to make utterance $U$ if it is informative, as long as it is true relative to the index

• listener makes joint inferences about indices and states of the world
**Lifted indices**

- Result: interpretation driven by expectations about AI’s strength.
- Statistical prior matters!
  - $u$ more informative as $\theta$ increases $\Rightarrow$ higher prob. speaker would say it.
  - But too-high $\theta$ means $u$ likely false.
- Predicted interpretation makes AI fairly strong, but not implausibly so.
Lifted indices

The graph shows the probability density functions of various parameters:

- **Threshold prior**
- **Threshold posterior**
- **Strength prior**
- **Strength posterior**

The x-axis represents the range from -6 to 6, and the y-axis represents the probability density from 0.0 to 0.4.
L₁ simulation: Upshot

- Context-sensitive probabilistic interpretation of RAs derived from
  - lexical meaning
  - background statistical knowledge about the relevant property
  - pragmatic preference for informative interpretations
- Borderline cases predicted
Absolute adjectives

- full/empty, wet/dry, safe/dangerous, ...
- meanings are less (not?) context-dependent
- meanings are sharp(er)
- reference classes apparently not relevant to interpretation
Absolute adjectives

- Sorites much less compelling
- A theater with all seats occupied is full.
- A theater with 1 less seat occupied than a full theater is also full.
- ∴ A theater with no seats occupied is full.
- Why?
Absolute adjectives

• Absolute interpretations occur when there is substantial prior mass on a scalar endpoint

• Relative interp. generated with little or no prior mass on scalar endpoints

• i.e., presence of endpoints allows but does not require relative interpretation

• Prediction: both possible with endpoints present, depending on form of prior
Simulation: full

Uniform prior on $\theta$
Simulation: full

Uniform prior on $\theta$

Uniform degree prior
Simulation: full

Uniform prior on $\theta$

Uniform degree prior

Inferred $\theta$ close to max
Simulation: **full**

Uniform prior on $\theta$

Uniform degree prior

Inferred $\theta$ close to max

Inferred degree very close to max
Simulation: full

Symmetric prior - symmetric predictions for max/max pair full/empty
Simulation: dangerous/safe

Uniform prior on $\theta$

Prior favors lower endpoint

Inferred $\theta$ close (but not equal) to 0

Inferred degree $\approx$ ‘greater than min’

safe not symmetric:
Inferred degree $\approx$ ‘at or near 0’
RAs on bounded scales

Uniform prior on $\theta$

Cost prior positive but small at 0

RA-like interpretation of expensive

non-minimal interpretation of cheap

![Graph showing density and cost priors]
AAs & RAs: Conclusions

• We don’t need special interpretive mechanisms or pragmatic principles to explain the absolute/relative distinction.

• Different interpretations generated by interaction btw scale structure, priors, pref. for informative interpretation.

• Changing prior can shift interpretation dramatically (full beer vs. wine glass)
The sorites

P1 A $10,000,000 house is expensive.

P2 A house $1 cheaper than an expensive house is also expensive.

P3 Therefore, a $1 house is expensive.
Sorites, v. I

Fixed-\(\theta\), universally quantified version

DP1  A \$10,000,000\ house is \(\theta\)-expensive.

DP2  For all houses \(x\) and \(y\): if \(x\) is \(\theta\)-expensive and \(y\) is \$1 cheaper than \(x\), then \(y\) is also \(\theta\)-expensive.

C  Therefore, a \$1\ house is \(\theta\)-expensive.

- Our semantics is bivalent: DP2 is false for any \(\theta\).
- Why is the sorites intuitively compelling?
- This isn’t how people understand the argument.
Sorites, v.2

sample $\theta$, $\text{cost}(x)$ given that $S_1$ says “$x$ is expensive”.

**SP1**  A $10,000,000$ house is $\theta$-expensive.  \[ P(\text{SP1}) \simeq 1 \]

**SP2**  If $x$ is $\theta$-expensive and $y$ is $1$ cheaper, then $y$ is also $\theta$-expensive.  \[ P(\text{SP2}) = ??? \]

**C**  Therefore, a $1$ house is $\theta$-expensive.  \[ P(\text{C}) \simeq 0 \]

- Sorites is compelling when $P(\text{SP2})$ is high.
- $P(\text{SP2})$ can be high without validating argument.
- Whether $P(\text{SP2})$ is high depends on joint posterior!
How high is $P(\text{SP2})$?

- Imagine a situation satisfying conditional antecedent: sample $\mu(x), A$ from $P_{L1}(\cdot | u)$
- Choose small gap $\varepsilon$ btw $\mu(x)$ and $\mu(y)$
- Check whether $\mu(y) > \theta$
How high is $P(\text{SP2})$?

- Sorites simulation results, with small $\varepsilon$
  - *tall*, Gaussian prior: $P(\text{SP2}) \approx .98$
  - *full*, uniform prior: $P(\text{SP2}) \approx .65$
  - *safe*, peaked prior: $P(\text{SP2}) \approx .7$

- Probability of inductive premise closely correlated with variance of $\theta$ posterior

- Prediction: vagueness, sorites susceptibility come in degrees
How high is $P(\text{SP2})$?

RAs: high variance $\Rightarrow$ high prob $\Rightarrow$ compelling

AAs: low variance $\Rightarrow$ low prob $\Rightarrow$ not compelling
Varying gap

![Graph showing varying gap with probability of inductive premise against \( \varepsilon \). The graph displays two lines, one for "tall" (blue) and one for "safe" (red).]
Previous work

• Previous probabilistic approaches to sorites simply assume some distribution $P(\theta)$:
  • Borel ’07, Black ’37, Kamp ’75, Edgington ’96, Frazee & Beaver ’11, Lassiter ’12

• This account is the first to explain
  • where $P(\theta)$ comes from
  • why the sorites doesn’t work with AAs
GAs: conclusions

• Simple semantics plus general Bayesian pragmatics predict reasonable probabilistic interpretations for GAs.

• All GAs have uncertain (vague) meanings to greater or lesser degrees.

• Different priors, constrained by scale structure, lead to different interpretations.

• Differences in sorites susceptibility emerge from scale structure, priors.
course summary

• probability and Bayesian inference
• noisy channels, audience design, modularity
• anchoring interpretation in world models
• Bayesian pragmatics
next steps

• lots of existing models we didn’t cover
• deepen connections with cognitive science
• explore predictions of linguistic theories
• lots of interesting linguistic, computational, philosophical issues to explore!
• be in touch if you want to get involved: danlassiter@stanford.edu
• Thanks!!
Appendix
AAs/RAs in definites

• Syrett, Kennedy & Lidz (2010): *The big glass* vs. *the empty glass*

• Very low probabilities with AAs leads to indep. difficulty in ps accommodation?

• Definites require clear differentiation of probabilities; probability that each option would count as ‘adj’ near-zero with *empty*, and so close
Multiple relative adjectives

• Puzzle: why don’t *warm* and *hot* mean the same thing (and *cool* and *cold*)?
  
• K07 seems to predict they should
  
• $x$ ‘stands out’ in warmness wrt $C$ iff $x$ stands out in hotness wrt $C$
  
• We can add to the model a constraint that $\theta$-hot > $\theta$-warm
  
• Reasonable vague interpretations emerge
Multiple relative adjectives

![Graphs showing the density of degree for different utterances](image)
Very

- Treat very A along similar lines:
  - A and very A compete
  - $\theta$-very-A > $\theta$-A
- Difference: added length of very A $\Rightarrow$ greater cost $\Rightarrow$ lower prior probability of utterance
Very

utt = "very short"

utt = "short"

utt = "tall"

utt = "very tall"