Disambiguation of Linguistic Uncertainty using Prior Epistemic Probabilities

JP Bernardy

January 23, 2023

Outline

RSA definition

$$P_{L_0}(w \mid u) \propto l(u, w) \times P(w) \tag{1}$$

$$P_{S_1}(u \mid w) \propto \frac{P_{L_0}(w \mid u)^{\alpha}}{e^{\alpha C(u)}} \tag{2}$$

$$P_{L_1}(w \mid u) \propto P_{S_1}(u \mid w) \times P(w) \tag{3}$$

Information-theoretic formulation of speaker model

$$G_{L_0}(u) = -\log \sum_{w \in \mathcal{W}} l(u, w) \times P(w) \tag{4}$$

$$P_{S_1}(u) \propto e^{\alpha(G_{L_0}(u) - C(u))} \tag{5}$$

$$P_{S_1}(u \mid w) \propto P_{S_1}(u) \times l(u, w) \tag{6}$$

Epistemic probability of utterances

Definition:

$$P_{L_0}(u) = e^{-G_{L_0}(u)}$$

Epistemic probability of utterances

Definition:

$$P_{L_0}(u) = e^{-G_{L_0}(u)}$$

Reformulation using epistemic probability:

$$P_{S_1}(u \mid w) \propto \frac{l(u, w)}{P_{L_0}(u)^{\alpha} \times e^{\alpha C(u)}} \tag{7}$$

3-factors, one for each Gricean maxim: Quality, Quantity, Economy

Example

- $t =$ "Al is tall"
- Semantics: $l(t, (h, \theta))$
  - $= 1$ if $h > \theta$
  - $= 0$ otherwise
- The meaning of "tall" has some ambiguity; in this example, the value of θ.
- Priors:
  - h normally distributed
  - θ uniform (To simplify, assume Lesbegue distribution)
- Can we disambiguate θ?
  - More precisely: get a posterior distribution for it given pragmatic effects.
- $C(t) = 2$

In general

- We call the uttered sentence t
- Any linguistic ambiguity attached to t can be represented by a random variable θ.
- Find posterior for (w, θ)

The silent alternative

Setup used in (some of) the literature:

- $\mathbf{S}$ utters t
- $\mathbf{L}$ considers the possibility that $\mathbf{S}$ would have remained silent.
- $\mathcal{U} = \{t, \emptyset\}$.
- $C(\emptyset) = 0$
- silence is (literally) compatible with every world: $\forall w.\ l(\emptyset, w) = 1$.

RSA instance for disambiguation

- The linguistic parameter is considered part of the situation to communicate:

$$w = (\theta, h)$$

- But $G_{L_0}(\emptyset) - C(\emptyset) = 0$.
- We deduce: $P_{S_1}(u) \propto f_{S_1}(u)$ with

$$f_{S_1}(t) = e^{\alpha(G_{L_0}(t) - C(t))} \quad (8)$$

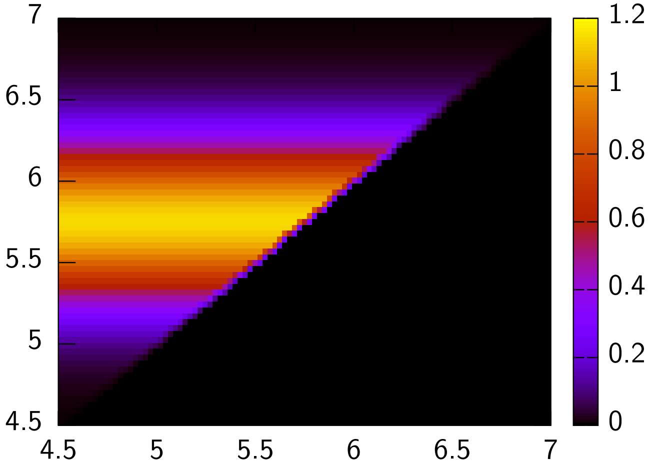
$$f_{S_1}(\emptyset) = e^{\alpha 0} = 1$$

- Therefore:

$$P_{S_1}(u = t) = \sigma(\alpha(G_{L_{0}(t)} - C(t))) \quad (9)$$

Example, Graphically

The effect of “Al is tall” on (h, θ) pairs (θ horizontal, h vertical):



Example, Conclusions

- Cost is 2 logits, or 2.89 bits
- Information gain is 1 bit.
- Silence wins (probabilistically)
  - If $\alpha = 4$, the probability of utterance of “Al is tall” is just 0.6 percent.

Speaker model, in general

$$P_{S_1}(u = t) = \sigma(\alpha(G_{L_0(t)} - C(t))) \tag{10}$$

Adding the dependency on w:

$$P_{S_1}(u = t \mid w) = l(t, w) \times \sigma(\alpha(G_{L_0(t)} - C(t))) \tag{11}$$

The speaker model has only two states, entirely determined by whether t is compatible with w. If incompatible, silence is the only option. If compatible, then the probability to utter t is a sigmoid function of its utility (gain minus cost) with temperature α.

Pragmatic listener, in general

$$\begin{aligned}
P_{L_1}(w \mid u = t) &\propto P_{S_1}(u = t \mid w) \times P(w) \\
&\propto l(t, w) \times \sigma(\alpha(G_{L_0(t)} - C(t))) \times P(w) && \text{by eq. (11)} \\
&\propto l(t, w) \times P(w) && (\dagger) \\
&\propto P_{L_0}(w \mid u = t) && \text{by def}
\end{aligned}$$

Conclusion

- We consider the whole space at once. The probability of utterance does not depend on θ, and thus there is no pragmatic effect.

Conclusion

- We consider the whole space at once. The probability of utterance does not depend on θ, and thus there is no pragmatic effect.
- RSA is pointless for semantic disambiguation.
- Give up?

Conclusion

- We consider the whole space at once. The probability of utterance does not depend on θ, and thus there is no pragmatic effect.
- RSA is pointless for semantic disambiguation.
- Give up? No!
- Modify RSA to get the desired outcome.

L-G model definition

$$P_{L_0}(w \mid u, \theta) \propto P(w) \times l(u, (w, \theta))$$

$$P_{S_1}(u \mid w, \theta) \propto \frac{P_{L_0}(w \mid u, \theta)^\alpha}{e^{\alpha C(u)}} \tag{12}$$

$$P_{L_1}(w, \theta \mid u) \propto P_{S_1}(u \mid w, \theta) \times P_{L_1}(w) \tag{13}$$

The L-G model: information-theoretic reformulation

Applying the treatment of the above section to L-G model, we get:

$$G_{L_0,\theta}(t) = -\log \sum_{w \in \mathcal{W}} l(t,(\theta,w)) \times P(w) \tag{14}$$

$$P_{S_1}(u = t \mid \theta) = \sigma(\alpha \times (G_{L_0,\theta}(t) - C(t))) \tag{15}$$

$$P_{S_1}(u = t \mid w, \theta) = l(t,(\theta,w)) \times P_{S_1}(u = t \mid \theta) \tag{16}$$

The L-G model: information-theoretic reformulation

Applying the treatment of the above section to L-G model, we get:

$$G_{L_0,\theta}(t) = -\log \sum_{w \in \mathcal{W}} l(t, (\theta, w)) \times P(w) \tag{14}$$

$$P_{S_1}(u = t \mid \theta) = \sigma(\alpha \times (G_{L_0,\theta}(t) - C(t))) \tag{15}$$

$$P_{S_1}(u = t \mid w, \theta) = l(t, (\theta, w)) \times P_{S_1}(u = t \mid \theta) \tag{16}$$

- The utility is now dependent on θ!

Example: when to utter; when to stay silent?

- It is more beneficial to utter "Al is tall" when the utility is positive:

$$G_{L_0,\theta}(isTall(\theta)) > C(isTall)$$

- Solve for θ
  - $\text{CDF}_{height}(\theta) \gtrsim 0.86$
  - θ one std. deviation over the normal (how convenient).
  - This value is entirely determined by the cost, $C(isTall)$
    - This cost is chosen by L-G, arbitrarily.

A model of impostor listeners

- The choice of utterance made by $\mathbf{S}$ is dependent on θ
  - e.g. the decision not to utter "Al is tall" if it's sufficiently obvious
- Choice can be made only if $\mathbf{S}$ already thinks that $\mathbf{L}$ knows the value of θ.
  - If $\mathbf{S}$ would think that $\mathbf{L}$ does not know— say $\mathbf{S}$ is trying to teach what "tall" means— then $\mathbf{S}$ would probably utter it as soon as it applies, as correctly predicted by the vanilla RSA model, above.
- The parametric variant of the RSA model corresponds to a scenario where
  - $\mathbf{L}$ has (a lot of) linguistic uncertainty
  - $\mathbf{L}$ believes that $\mathbf{S}$ believes that there is none.
- L-G model a situation where $\mathbf{L}$ is learning the language, but appears to $\mathbf{S}$ as if it would already know it.
  - This is an "incognito ignorant" or "impostor" listener model

General pragmatic listener model

Apply the same recipe as for the vanilla model, to get:

$$P_{L_1}(w, \theta \mid u = t) \propto l(t, (\theta, w)) \times \sigma(\alpha(G_{L_0,\theta}(t) - C(t))) \times P(w)$$

Pragmatically guessing θ

- Marginalize away w; focus on θ.
- Take the average over w

$$P_{L_1}(w, \theta \mid u = t) \propto l(t, (\theta, w)) \times \sigma(\alpha(G_{L_0,\theta}(t) - C(t))) \times P(w)$$

$$\begin{aligned} P_{L_1}(\theta \mid u = t) &\propto \left(\sum_{w \in \mathcal{W}} P(w) \times l(t, (\theta, w)) \right) \times \sigma(\alpha(G_{L_0,\theta}(t) - C(t))) \\ &\propto \exp(-G_{L_0,\theta}(t)) \times \sigma(\alpha(G_{L_0,\theta}(t) - C(t))) \end{aligned}$$

- The utterance t affects the posterior distribution of θ *only* through its cost and the literal information/epistemic probability associated with $t(\theta)$.

$$P_{L_1}(\theta \mid u = t) \propto \frac{p}{1 + \left(\frac{p}{\gamma}\right)^{\alpha}}$$

  - with $p = P_{L_0}(t(\theta))$ and $\gamma = \exp(-C(t))$

Setting up the paradigm shift

Final reformulation of L-G model

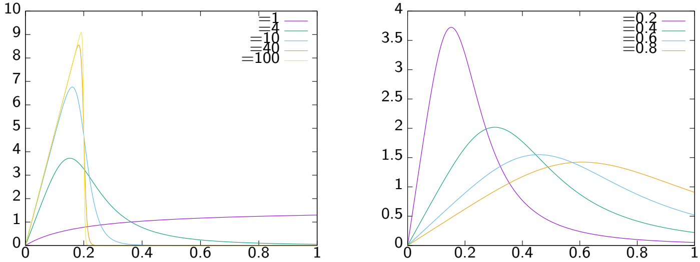
- The epistemic probability associated with the interpretation $t(\theta)$ follows the $\mathrm{SharkFin}$ distribution.

$$P_{L_0}(t(\theta)) \sim \mathrm{SharkFin}(\alpha, \gamma) \tag{17}$$

with

$$\mathrm{SharkFin}(\alpha, \gamma; p) \propto \frac{p}{1 + \left(\frac{p}{\gamma}\right)^{\alpha}}$$

The sharkfin distribution



(a) Varying α, with $\gamma = 0.2$

(b) Varying γ, with $\alpha = 4$

Figure: The SharkFin distribution for various values of its parameters.

The take home message

- The L-G models tells us that the listener expects the epistemic probability associated with interpretation θ to follow the $\mathrm{SharkFin}$distribution.
- ie. The above equation gives us a way to disambiguate θ on the basis of if the listener believes (a priori) $t(\theta)$
- CLAIM: there is nothing special about SharkFin!

ISA

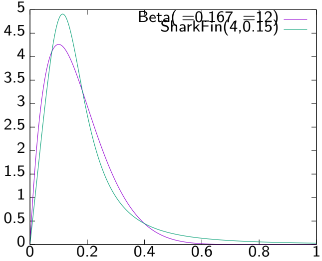
To interpret an ambiguous utterance charitably, apply the following Bayesian update on the distributions of interpretations.

$$\boxed{P_{L_0}(t(\theta)) \sim G} \tag{18}$$

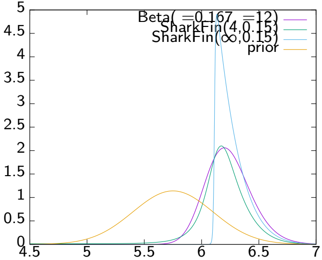
where G:

- Is a continuous probability distribution on the interval $[0, 1]$
- $\text{PDF}_G(0) = 0$; corresponding to the Maxim of Quality (impossible interpretations are rejected)
- $\text{PDF}_G(1) = 0$; corresponding to the Maxim of Quantity (uninformative interpretations are rejected)
- Monomodal
- For the rest, we should choose G on the basis of experimental evidence.

G from Beta family



(a) Beta and the SharkFin distributions.



(b) θ posterior, for h normally distributed and various predicating distributions (G).

Figure: Approximating the behaviour of Goodman-Lassiter RSA by a Beta distribution predication on epistemic probabilities. The baseline corresponds to the $\mathrm{SharkFin}$ distribution for the parameters chosen by L-G.

ISA consequences

- Speakers are assumed to try an communicate some information (G quantifies how much)
- When uttering a sentence, a speaker may be ambiguous. They can assume that the listener will be charitable and select a reasonable interpretation (according to the previous point).

Posteriors for vanilla rsa, Example.

- No pragmatic effects. So we use the literal probabilities.

$$\begin{aligned} P(h,\theta|t) &\propto P(h)\mathbb{1}(h>\theta) \\ &= \frac{P(h)\mathbb{1}(h>\theta)}{\int dh \int d\theta P(h)\mathbb{1}(h>\theta)} \\ &= 2P(h)\mathbb{1}(h>\theta) \end{aligned}$$

Posteriors for vanilla rsa, Example ( ).

Marginalizing. To simplify the expressions we imagine the domain is infinite.

$$\begin{aligned} P(\theta|t) &\propto \int_{-\infty}^{+\infty} dh 2P(h)\mathbb{1}(h > \theta) \\ &\propto \int_{\theta}^{+\infty} P(h) \\ &\propto 1 - CDF_{height}(\theta) \end{aligned}$$

(θ goes from uniform to skewed to the left of the median height.)

Posteriors for vanilla rsa, Example (height).

$$
\begin{aligned}
P(h|t) &\propto \int_{-\infty}^{+\infty} d\theta 2P(h)\mathbb{1}(h > \theta) \\
&\propto P(h)\int_{-\infty}^{h} d\theta \\
&\propto P(h)CDF_{threshold}(h)
\end{aligned}
$$