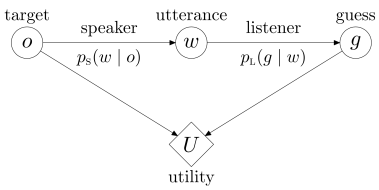
A Game-Theoretic Approach to Generating Spatial Descriptions

Golland, Liang & Klein (2010)

- **Introduction**
  - Many semantically valid utterances, far fewer pragmatically licensed (goal-achieving) utterances
    (a) *right of* `O2` (b) *on* `O3`
  - Enter game theory: speaker and listener can share the same utility function

The Communication Game

1. In order to communicate a *target* o to L, S produces an *utterance* w chosen according to a strategy $p_S(w|o)$.
2. L interprets w, responds with *guess* g according to $p_L(g|w)$.
3. S and L collectively get a utility of $U(o,g) \stackrel{\text{def}}{=} \mathbb{I}[o=g]$.



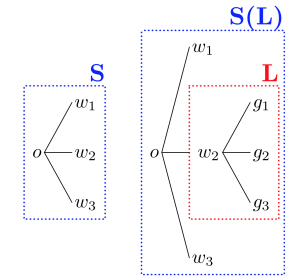
$$\text{Expected utility EU(S,L)} = \sum_{o,w,g} p(o)p_S(w|o)p_L(g|w)U(o,g)$$
$$= \sum_{o,w} p(o)p_S(w|o)p_L(o|w)$$

- S:LITERAL selects (a) or (b) each with probability $\frac{1}{2}$
  - (a): L:LITERAL guesses correctly with probability $\frac{1}{2}$; (b): guesses correctly with prob. 1
- EU(S:LITERAL, L:LITERAL)= $\frac{3}{4}$

- **From reflex speaker to rational speaker**
  - Rational speakers optimize their expected utility using a listener model:

$$p_{S(L)}(w|o) = \mathbb{I}[w = w^*], \text{where}$$
$$w^* = \underset{w'}{\text{argmax}}\ p_L(o|w')$$

  - $p_{L:\text{LITERAL}}(\texttt{O1}|\text{a} = \frac{1}{2})$
  - $p_{L:\text{LITERAL}}(\texttt{O1}|\text{b} = 1)$



(a) Reflex speaker (b) Rational speaker

- **From literal speaker to learned speaker**
- How can we improve literal strategies with learning?
- Experiments:
  * Speakers prompted with a target object o and asked to produce an utterance w
  * Listeners given an utterance w and asked to guess object o
- Trained a log-linear speaker/listener

Review of log-linear models

- Model distributions of the form $P(Y|X)$, where Y ranges over a countable set of response classes y_i
  - e.g. utterances w
- FEATURE FUNCTIONS $f_j(X, Y)$ map every possible paired instance of X and Y to a real number
  - e.g. distance between target object and reference object
- Each possible response class y_i has a feature vector
  - e.g. each utterance w has a feature vector $\phi(o, w)$
- Each feature function f_j has a w has a corresponding parameter λ_j
  - e.g. θ_{S}
- The conditional probability of each class y_i is defined to be:

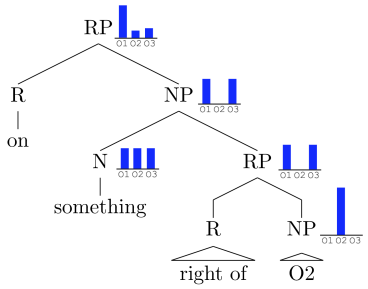
$$P(Y = y_i|X = x) = \frac{1}{Z}\exp\left[\sum_j \lambda_j f_j(x, y_i)\right]$$

$$p_{\text{S: LEARNED}}(w|o; \theta_{\text{S}}) \propto \exp\{\theta_{\text{S}}^\top \phi(o, w)\}$$
$$p_{\text{S: LEARNED}}(g|w; \theta_{\text{L}}) \propto \exp\{\theta_{\text{L}}^\top \phi(w, w)\}$$

- Speaker and listener use the same set of features, but they have different parameters
- Features: proximity functions, topological functions, projection functions

- **Handling complex utterances**

[noun]	N	$\rightarrow$	*something* | O1 | O2 | $\cdots$
[relation]	R	$\rightarrow$	*in front of* | on | $\cdots$
[conjunction]	NP	$\rightarrow$	N RP*
[relativization]	RP	$\rightarrow$	R NP



  - Each node in a parse tree has a denotation $[\![w]\!]$, a distribution over objects in the scene
  - For a subtree w

$$p_{\text{L}}(g|w) \propto \begin{cases} \mathbb{I}[g \in \mathcal{N}(x)] & w = (\text{N } x) \\ \prod_{j=1}^{k} p_{\text{L}}(g|w_j) & w = (\text{NP } w_1 \ldots w_k) \\ \sum_{g'} p_{\text{L}}(g|(r, g'))p_{\text{L}}(g'|w') & w = (\text{RP R } w') \end{cases}$$

x = single child of w and $\mathcal{N}(x)$=objects consistent with x

g' = objects in the child NP tree

- **Modeling listener confusion**
  - Let $\alpha \in [0, 1]$ be a focus parameter which determines the confusion level

$$\tilde{p}_{\text{L}}(g|w) = \alpha^{|w|} p_{\text{L}}(g|w) + (1 - \alpha^{|w|}) p_{\text{rnd}}(g|w)$$

  - As $\alpha \rightarrow 0$, the confused listener is more likely to make a random guess, and thus there is a stronger penalty against using more complex utterances

- **Evaluation**

– Utility: average the utility (communicative success) over the test scenarios Ts

$$\begin{aligned}\textsc{Success}(\textsc{s}) &= \textsc{EU}(\textsc{s},\textsc{l}{:}\textsc{Human}) \\ &= \frac{1}{|\textsc{Ts}|}\sum_{o\in\textsc{Ts}}\sum_{w} p_{\textsc{s}}(w|o)p_{\textsc{l}:\ \textsc{Human}}(o|w)\end{aligned}$$

– Exact match: ability of their speaker to exactly match an utterance produced by a human speaker

$$\textsc{match}(\textsc{s}) = \frac{1}{|\textsc{Ts}|}\sum_{o\in\textsc{Ts}}\sum_{w} p_{\textsc{s}:\ \textsc{Human}}(w|o)p_{\textsc{s}}(w|o)$$

Speaker	Success	Exact Match
s:Literal [reflex]	4.62%	1.11%
s(l:Literal) [rational]	33.65%	2.91%
s:Learned [reflex]	38.36%	5.44%
s(l:Learned) [rational]	**52.63%**	14.03%
s:Human	41.41%	**19.95%**

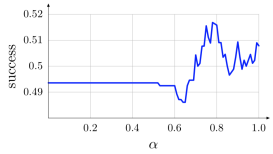


Figure 10: Communicative success as a function of focus parameter α without tabooing on TsDev. The optimal value of α is obtained at 0.79.

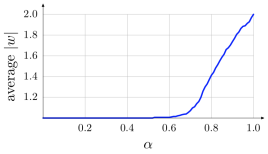


Figure 11: Average utterance complexity as a function of the focus parameter α on TsDev. Higher values of α yield more complex utterances.

Discussion/commentary

- Why are the log-linear model parameters allowed to differ between speaker and listener? What, if any, systematic differences would we expect?
- Should the focus parameter capture linguistic complexity, or propositional complexity? We can convey complex ideas using *conjunctions of* simple structures: *Go into the kitchen and you'll see a cabinet above the refrigerator. On the right side of the cabinet...*
- What about an utterance *length* parameter (corresponding to the conjunction rule)? Here we might expect the opposite behavior for α: the more information the speaker provides, the less confused the listener should be
- The rational model does pretty terribly at matching human utterances exactly, but winds up with higher utility (communicative success). What do we make of this?