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

<table>
<thead>
<tr>
<th>The Communication Game</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. In order to communicate a target o to l, s produces an utterance w chosen according to a strategy ps(w</td>
</tr>
<tr>
<td>2. l interprets w, responds with guess g according to pl(g</td>
</tr>
<tr>
<td>3. s and l collectively get a utility of U(o,g) def = I[o = g].</td>
</tr>
</tbody>
</table>

Expected utility EU(s,l) = ∑∑∑ p(o)p_s(w|o)p_l(g|w)U(o,g)
= ∑∑ p(o)p_s(w|o)p_l(o|w)

– s:LITERAL selects (a) or (b) each with probability 1/2
  – (a): l:LITERAL guesses correctly with probability 1/2; (b): guesses correctly with prob. 1
  – EU(s:LITERAL, l:LITERAL) = 3/4

• From reflex speaker to rational speaker
  – Rational speakers optimize their expected utility using a listener model:
    ps(l)(w|o) = I[w = w*], where
    w* = argmax_{w'} p_l(o|w')

  – pl:LITERAL(01|a = 1/2)
  – pl:LITERAL(01|b = 1)

• 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 \( \lambda_j \) corresponding parameter
  
  - e.g. \( \theta_s \)

- The conditional probability of each class \( y_i \) is defined to be:
  
  \[
P(Y = y_i|X = x) = \frac{1}{Z} \exp \left[ \sum_j \lambda_j f_j(x, y_i) \right]
  \]

\[
p_{s: \text{learned}}(w|o; \theta_s) \propto \exp\{\theta_s^T \phi(o,w)\}
\]

\[
p_{s: \text{learned}}(g|w; \theta_l) \propto \exp\{\theta_l^T \phi(w,w)\}
\]

- Speaker and listener use the same set of features, but they have different parameters

- Features: proximity functions, topological functions, projection functions

- **Handling complex utterances**

- Each node in a parse tree has a denotation \([w]\), a distribution over objects in the scene

- For a subtree \( w \)

  \[
p_l(g|w) \propto \left\{ \begin{array}{ll}
  \mathbb{I}[g \in N(x)] & w = (N x) \\
  \prod_{j=1}^{k} p_r(g|w_j) & w = (NP \ w_1 \ldots \ w_k) \\
  \sum_{g'} p_r(g|(r, g')) p_l(g'|w') & w = (RP \ r \ w') \\
  \end{array} \right.
  \]

  \( g' \) = objects in the child NP tree

- **Modeling listener confusion**

  - Let \( \alpha \in [0,1] \) be a focus parameter which determines the confusion level

    \[
    \tilde{p}_l(g|w) = \alpha^{|w|} p_l(g|w) + (1 - \alpha^{|w|}) p_{\text{rand}}(g|w)
    \]

  - As \( \alpha \to 0 \), 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

\[
\text{SUCCESS}(s) = \text{EU}(s,l:\text{HUMAN}) = \frac{1}{|Ts|} \sum_{o \in Ts} \sum_{w} p_s(w|o)p_l: \text{HUMAN}(o|w)
\]

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

\[
\text{MATCH}(s) = \frac{1}{|Ts|} \sum_{o \in Ts} \sum_{w} p_s: \text{HUMAN}(w|o)p_s(w|o)
\]

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Success</th>
<th>Exact Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.(LITERAL) [reflex]</td>
<td>4.62%</td>
<td>1.11%</td>
</tr>
<tr>
<td>S.(LITERAL) [rational]</td>
<td>33.65%</td>
<td>2.91%</td>
</tr>
<tr>
<td>S.(LEARNED) [reflex]</td>
<td>38.36%</td>
<td>5.44%</td>
</tr>
<tr>
<td>S.(LEARNED) [rational]</td>
<td>52.63%</td>
<td>14.03%</td>
</tr>
<tr>
<td>S.HUMAN</td>
<td>41.41%</td>
<td>19.95%</td>
</tr>
</tbody>
</table>

Figure 10: Communicative success as a function of focus parameter $\alpha$ without tweaking on TsDEV. The optimal value of $\alpha$ is obtained at 0.79.

Figure 11: Average utterance complexity as a function of the focus parameter $\alpha$ on TsDEV. Higher values of $\alpha$ 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 $\alpha$: 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?