Rational Analysis

• “An algorithm is likely understood more readily by understanding the **nature of the problem** being solved than by examining the mechanism (and the hardware) in which it is solved.” (Marr, p27)

• **Principle of Rationality:** The cognitive system optimizes the adaptation of the behavior of the organism.

• If a cognitive processes is viewed as rational, then the **computational** theory should reflect optimal adaptation to the task & environment:

  1. Derive the Optimal Function
  2. Test against the empirical data
  3. Revise the Optimal Function
Garden Path vs. Garden Variety

• Human Language Processing: **Garden Paths**
  
  ✗ Incremental disambiguation process can fail
  
  ✗ Memory limitations lead to breakdown
  
  ✗ Garden paths lead to misinterpretations, complexity or breakdown

• Human Language Processing: **Garden Variety**
  
  ✔ Accurate: typically recover the correct interpretation
  
  ✔ Robust: are able to interpret ungrammatical & noisy input
  
  ✔ Fast: people process utterances in real-time, incrementally

Can we treat language as a rational cognitive system?

Relating Models with Data

Computational

Algorithmic

Implementational

Optimal function for the task?

Resource limitations

Empirical Data
arg\max_i P(s_i) \text{ for all } s_i \in S

- Empirical: lexical access, word category/sense, subcategorization
- Rational: accurate, robust, broad coverage
- Rational Models:
  - explain accurate performance in general: i.e. rational behaviour
  - explain specific observed human behavior: e.g. for specific phenomena

## Lexical Category Disambiguation

- Sentence processing involves the resolution of lexical, syntactic, and semantic ambiguity.
  - Solution 1: These are not distinct problems
  - Solution 2: Modularity, divide and conquer
- Category ambiguity:
  - *Time flies like an arrow.*
- Extent of ambiguity:
  - 10.9% (types) 65.8% (tokens) (Brown Corpus)
The Model: A Simple POS Tagger

- Find the best category path \((t_1 \ldots t_n)\) for an input sequence of words \((w_1 \ldots w_n)\): \(P(t_0, \ldots t_n, w_0, \ldots w_n)\)

- Initially preferred category depends on two parameters:
  - Lexical bias: \(P(w_i|t_i)\)
  - Category context: \(P(t_i|t_{i-1})\)

- Categories are assigned incrementally: Best path may require revision

### 2 Predictions

- The Statistical Hypothesis:
  - Lexical word-category frequencies, \(P(w_i|t_i)\), are used for initial category resolution

- The Modularity Hypothesis:
  - Initial category disambiguation is modular, and not determined by (e.g. syntactic) context beyond \(P(t_i|t_{i-1})\).

- Two experiments investigate
  - The use word-category statistics
  - Autonomy from syntactic context
Statistical Lexical Category Disambiguation

• Initially preferred category depends on: \( P(t_0, \ldots, t_n, w_0, \ldots, w_n) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}) \)

• Categories are assigned incrementally
  • the warehouse *prices* the beer very modestly
    DET N N/V VI
  • the warehouse *prices* are cheaper than the rest
    DET N N/V N ...
  • the warehouse *makes* the beer very carefully
    DET N N/V V ...
  • the warehouse *makes* are cheaper than the rest
    DET N N/V N! ...

• Interaction between bias and disambiguation
• Category frequency determines initial decisions

Lexical bias: \( P(w_i|t_i) \)
Category context: \( P(t_i|t_{i-1}) \) – constant!
Trained on the Susanne corpus

Modular Disambiguation?

• Do initial decisions reflect integrated use of both lexical and syntactic constraints/biases or just (modular) lexical category biases?
  • N/V bias with immediate/late syntactic disambiguation as noun

• Main effect of bias at disambiguation:
  • Initial decisions ignore syntactic context.
  • Problematic for lexicalist syntactic theories
  • At c2, VA/VU difference is significant
  • Implies lexical category doesn’t include number (?!)

a) [V-bias, N-disamb] The warehouse *makes* are cheaper than the rest.
b) [V-bias, N-unamb] The warehouse *make* is cheaper than the rest.
c) [N-bias, N-disamb] The warehouse *prices* are cheaper than the rest.
d) [N-bias, N-unamb] The warehouse *price* is cheaper than the rest.
‘That’ Ambiguity \quad \text{(Juliano & Tanenhaus)}

A. \textit{That} \textit{experienced diplomat(s)} would be very helpful ... \,[DET]

B. The lawyer insisted \textit{that experienced diplomat(s)} would be very helpful \,[Comp]

- Initially: \(\text{det}=.35\quad \text{comp}=.11\)
- Post-verbally: \(\text{comp}= .93\quad \text{det}= .06\)

- Found increased RT when dispreferred (according to context) is forced in the disambiguation region “diplomat(s)”

- Advocates bigram over unigram:

  \[
  \begin{array}{|c|c|c|}
  \hline
  t_i & \text{Comp} & \text{Det} \\
  \hline
  t_{i-1} = \text{verb} & .0234 & .0051 \\
  t_{i-1} = \text{start} & .0003 & .0111 \\
  \hline
  \end{array}
  \]

  \[
  \begin{align*}
  P(\text{that}|\text{comp}) &= 1, \quad P(\text{that}|\text{det}) = .171 \\
  P(\text{comp}|\text{verb}) &= .0234, \quad P(\text{det}|\text{verb}) = .0296 \\
  P(\text{comp}|\text{start}) &= .0003, \quad P(\text{det}|\text{start}) = .0652 \\
  \end{align*}
  \]

\textbf{Internal Reanalysis}

- The tagger model predicts internal reanalysis for some sequences.

- Viterbi: revise most likely category sequence based on new evidence

- Right context in RR/MV ambiguities: [MacDonald 1994]

  - The sleek greyhound \textit{raced at the track} won the event
  
  - The sleek greyhound \textit{admired at the track} won the event

- \textit{raced} = intrans bias, \textit{admired} = trans bias

- Increased RT \,(blue)\ indicate transitivity bias is used
An SLCM Account

• Assume transitive/intransitive POS categories, extract frequencies from the Susanne corpus:

The man fought at the police station fainted
[intransitive]
The man held at the police station fainted  [transitive]

Predicts garden path for intransitives

Predicts rapid reanalysis for transitives

Reduced Relative Clause

• Parsers can make wrong decisions that lead them up the garden path

“The man raced to the station was innocent”

The Problem

• In some cases it may be possible to recover from the error earlier

“The man held at the station was innocent”


SLCM Summary

• Psychologically plausible: lower statistical complexity than other models

• High accuracy in general: explains why people perform well overall

• Explains where people have difficulty
  • Statistical: category frequency drives initial category decisions
  • Modular: syntax structure doesn’t determine initial category decisions
  • Bigram evidence: “that” ambiguity [Juliano and Tanenhaus]
  • Reanalysis of verb transitivity for ‘reduced relatives’ [MacDonald]
A Puzzle

• Sometimes local thematic assignment appears to violate global parse:
  • [A/R] The coach smiled at the player tossed a frisbee by the ...
  • [U/R] The coach smiled at the player thrown a frisbee by the ...
  • [A/U] The coach smiled at the player who was tossed a frisbee by the ...
  • [U/U] The coach smiled at the player who was thrown a frisbee by the ...

• Do people consider the locally coherent, but not globally licensed parse?
  • This structure shouldn’t even be considered by incremental parsers

Results:

• These results are problematic for theories requiring global contextual consistency (e.g. Frazier, 1987; Gibson, 1991, 1998)
An SLCM Account

- Initially preferred category depends on two parameters:
  - Lexical bias: \( P(w_i|t_i) \)  Category context: \( P(t_i|t_{i-1}) \)

- [AS-AV] The coach smiled at the player *tossed* a frisbee

  \[
P(\text{tossed}|\text{Vpast}) \times P(\text{Vpast}|\text{noun}) > P(\text{tossed}|\text{Vpart}) \times P(\text{Vpart}|\text{noun})
\]

  So: assign *tossed=Vpast*, but can’t integrate into parse, so reanalyse

- [US-AV] The coach smiled at the player who was *tossed* a frisbee

  \[
P(\text{tossed}|\text{Vpast}) \times P(\text{Vpast}|\text{Aux}) < P(\text{tossed}|\text{Vpart}) \times P(\text{part}|\text{Aux})
\]

  So: assign *tossed=Vpastparse*, integrate into parse, no difficulty

Comments on the SLCM

- Evidence category preference appears truly frequency-based
- Indication of which features are exploited [e.g. transitivity, not number]
  - But this is subject to further empirical investigation & verification
- Combines optimality of probabilities with advantages of modularity
  - psychological plausibility due to tractable parameter space
- Implications for the **Grain Problem**?
  - Bigrams used, but not tri-grams, or syntactic structure ?
  - Transitivity but not number ? More/less syntactically-rich POS tags ?
Probabilistic Syntax

• The SLCM is only a model of lexical category assignment
  • But note: these category decisions underlie many “syntactic” ambiguities
• Some ambiguities are purely syntactic, however:
  • Relative clause attachment, or other modifier attachment
  • NP/S complement ambiguity (unless subcat is encoded in the POS tags)
• Also evidence that compositional interpretation influences parsing
  • Can’t be modeled in the SLCM alone
• Apply probabilistic approaches to modeling human syntactic parsing

Probabilistic Language Processing

• Task of comprehension: recover the correct interpretation
• Goal: Determine the most likely analysis for a given input:
  \[
  \arg \max_i P(s_i) \text{ for all } s_i \in S
  \]
• P hides a multitude of sins:
  • P corresponds to the degree of belief in a particular interpretation
  • Influenced by recent utterances, experience, non-linguistic context
• P is usually determined by frequencies in corpora or completions
• To compare probabilities (of the Si), we assume parallelism. How much?
Estimating P: The Grain Problem

• Suppose you have been exposed to N sentences in your lifetime

• “Our company is training workers”

• Problem: P=0, often

• Solution: Estimate P, by combining probabilities of smaller chunks

\[
P(S=s1) = \frac{C(s1)}{N} \]
\[
P(S=s2) = \frac{C(s2)}{N} \]
\[
P(S=s3) = \frac{C(s3)}{N} \]

PCFGs: a quick reminder

• Context-free rules annotated with probabilities

• Probabilities of all rules with the same LHS sum to one;

• Probability of a parse is the product of the probabilities of all rules applied.

• Example (Manning and Schütze 1999)

<table>
<thead>
<tr>
<th>Rule</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → NP VP</td>
<td>1.0</td>
</tr>
<tr>
<td>PP → P NP</td>
<td>1.0</td>
</tr>
<tr>
<td>VP → VP NP</td>
<td>0.7</td>
</tr>
<tr>
<td>VP → VP PP</td>
<td>0.3</td>
</tr>
<tr>
<td>P → with</td>
<td>1.0</td>
</tr>
<tr>
<td>V → saw</td>
<td>1.0</td>
</tr>
<tr>
<td>NP → NP PP</td>
<td>0.4</td>
</tr>
<tr>
<td>NP → astronomers</td>
<td>0.1</td>
</tr>
<tr>
<td>NP → ears</td>
<td>0.18</td>
</tr>
<tr>
<td>NP → saw</td>
<td>0.04</td>
</tr>
<tr>
<td>NP → stars</td>
<td>0.18</td>
</tr>
<tr>
<td>NP → telescopes</td>
<td>0.1</td>
</tr>
</tbody>
</table>
The Grain Problem

- Experience-based models rely on frequency of prior linguistic exposure to determine preferences. What kinds of things do we count?
  - Complete sentence/structure occurrences? Data too sparse.
  - Lexical: Verb subcategorization frequencies. Should we distinguish tenses? Senses?
  - Word level: specific word forms or lemmas? Part-of-speech, how detailed?
  - Tuning is structural: NP P NP RC vs NP P NP RC
    
    
    
    High
    Low

- Does all experience have equal weight (old vs. new)?

- Are more frequent “words” or “strings” (idioms) dealt with using finer grain statistics than rarer expressions?

Relative Clause Attachment

\[
\text{S} \rightarrow \text{NP} \rightarrow \text{VP} \\
\text{NP} \rightarrow \text{V} \rightarrow \text{PP} \\
\text{PP} \rightarrow \text{NP} \\
\text{RC}
\]

\[
\text{Someone} \rightarrow \text{V} \rightarrow \text{the servant} \rightarrow \text{of} \rightarrow \text{the actress} \rightarrow \text{who was on the balcony}
\]

\[
\text{Alguien disparo contra el criado de la actriz que estaba en el balcón}
\]