Probabilistic Syntax Processing

• Lexical frequencies can contribute to resolving many ambiguities, but not all.

• Does human parser keep track of structural as well as lexical frequencies?

• Sometimes in contrast with previously suggested principles, such as Late Closure (Frazier)

  Someone shot the servant of the actress who was on the balcony.
Relative Clause Attachment

Alguien disparo contra el criado de la actriz que estaba en el balcón

Someone V NP
shot NP PP
the servant of NP RC
the actress who was on the balcony
### Cross-linguistic RC Preferences

<table>
<thead>
<tr>
<th>Language</th>
<th>Off-line</th>
<th>On-line</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>French</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Italian</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Dutch</td>
<td>high</td>
<td>low(early), high(late)</td>
</tr>
<tr>
<td>German</td>
<td>high</td>
<td>low(early), high(late)</td>
</tr>
<tr>
<td>English</td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>Arabic</td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>Norwegian</td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>Swedish</td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>Romanian</td>
<td>low</td>
<td>low</td>
</tr>
</tbody>
</table>

- Experienced-based treatment of structural ambiguity?
Tuning Hypothesis

• **Tuning Hypothesis** (Mitchell et al., 1995):
  
  • Human parser deals with ambiguity by initially selecting the syntactic analysis that has worked most frequently in the past.
  
  • Further evidence: school children’s preferences before and after a period of two weeks in which exposure to high/low examples was increased (Cuetos et al., 1996)

• How to formalize this hypothesis?
The Competition Model

- **The Competition Model** (MacWhinney et al. 1984)

- **Goal**: map from the formal level (surface forms, syntactic constructions, etc) to functional level (meaning, intention)

- **Approach**: probabilistically combine various surface cues for choosing the correct functional interpretation

- Focus on the combination of cues, and how the probabilities vary from language to language

- E.g., assigning thematic roles to grammatical positions (English: word order; German: morphological cues)
Cue Validity

• Cue validity $v(c,i)$: contribution of a cue $c$ to an interpretation $i$
  
  $v(c,i) = \text{availability}(c) \times \text{reliability}(c,i)$

  
  $P(c) \times P(i|c) = P(c,i)$

• Combining various cues: $\prod_i P(A|c_i)$

• Comparing two interpretations $A$ and $B$:

  
  $P(A|C) = \frac{\prod_i P(A|c_i)}{\prod_i P(A|c_i) + \prod_i P(B|c_i)}$
Probabilistic Parsing

• Considering the N sentences seen in the past, choose the structure with the highest probability

How to calculate the probability of a sentence?
• Maximum likelihood estimation: \( P(S) = \frac{C(S)}{N} \)
• **Grain problem**: \( C(S1) = C(S2) = 0 \); better use probabilities of the smaller chunks, but how small?
### Stochastic CFGs

- Augment standard context free grammars by annotating grammar rules with probabilities.

<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 NP</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>

- 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.
Parse Ranking

$t_1$:

```
S_{1.0}  
   |                  
NP_{0.1}  VP_{0.7}  
  |                  
astronomers      NP_{0.4}  
                |                  
V_{1.0}  PP_{1.0}  
  |                  
saw      NP_{0.18}  
          |                  
stars    PP_{1.0}  
          |                  
with     NP_{0.18}  
          |                  
ears
```

$P(t_1) = 1.0 \times 0.1 \times 0.7 \times 1.0 \times 0.4 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0009072$
\[ P(t_1) = 1.0 \times 0.1 \times 0.3 \times 0.7 \times 1.0 \times 0.18 \times 1.0 \times 1.0 \times 0.18 = 0.0006804 \]
Jurafsky (1996)

• Psycholinguistic model of lexical and syntactic access and disambiguation

• Probability of a parse is a combination of
  • Stochastic CFGs
  • Frame probabilities of individual items

• Architecture: incremental, bounded parallel
  • Computation of parse probabilities is incremental
  • Least probable parses are pruned
The women discussed the dogs on the beach.

*t1*: The women discussed them (the dogs) while on the beach.

*t2*: The women discussed the dogs which were on the beach.

\[ p(\text{discuss}, \langle \text{NP PP} \rangle) = 0.24 \]
\[ p(\text{discuss}, \langle \text{NP} \rangle) = 0.76 \]
\[ \text{VP} \rightarrow \text{V NP XP} \quad 0.15 \]
\[ \text{VP} \rightarrow \text{V NP} \quad 0.39 \]
\[ \text{NP} \rightarrow \text{NP XP} \quad 0.14 \]

\[ p(t_1) = 0.15 \times 0.24 = 0.036 \text{ (dispreferred)} \]
\[ p(t_2) = 0.76 \times 0.39 \times 0.14 = 0.041 \text{ (preferred)} \]
The women kept the dogs on the beach.

✓ $t_1$: The women kept them (the dogs) on the beach.

$t_2$: The women kept the dogs which were on the beach.

$p(\text{keep}, \langle \text{NP XP[pred +]} \rangle) = 0.81$

$\text{VP} \rightarrow \text{V NP XP} \quad 0.15$

$t_1$:

\[
\begin{array}{c}
\text{VP} \\
\quad \text{V} \\
\quad \quad \text{NP} \\
\quad \quad \quad \text{keep} \\
\quad \quad \quad \text{the dogs} \\
\quad \quad \quad \text{on the beach} \\
\end{array}
\]

$p(t_1) = 0.15 \times 0.81 = 0.12$ (preferred)

$p(\text{keep}, \langle \text{NP} \rangle) = 0.19$

$\text{VP} \rightarrow \text{V NP} \quad 0.39$

$\text{NP} \rightarrow \text{NP XP} \quad 0.14$

$t_2$:

\[
\begin{array}{c}
\text{VP} \\
\quad \text{V} \\
\quad \quad \text{NP} \\
\quad \quad \quad \text{keep} \\
\quad \quad \quad \text{the dogs} \\
\quad \quad \quad \text{on the beach} \\
\end{array}
\]

$p(t_2) = 0.19 \times 0.39 \times 0.14 = 0.01$ (dispreferred)
Construction Preferences

S → NP ... 0.92
NP → Det Adj N 0.28
N → ROOT s 0.23
N → house 0.0024
Adj → complex 0.00086

$t_1$: 

\[ p(t_1) = 1.2 \times 10^{-7} \text{ (preferred)} \]

NP → Det N 0.63
S → [NP VP[V ... 0.48
N → complex 0.000029
V → house 0.0006
V → ROOT s 0.086

$t_1$: 

\[ p(t_1) = 4.5 \times 10^{-10} \text{ (dispreferred)} \]
Construction Preferences

$S \rightarrow \text{NP} \ldots \quad 0.92$
$\text{NP} \rightarrow \text{Det} \quad N \quad N \quad 0.28$
$N \rightarrow \text{fire} \quad 0.00072$
$N \rightarrow \text{ROOT s} \quad 0.23$

\[ t_1: \]

\[
\begin{array}{c}
\text{S} \\
\downarrow \\
\text{NP} \\
\downarrow \\
\text{Det} \\
\quad \text{the} \\
\downarrow \\
\text{N} \\
\quad \text{warehouse} \\
\downarrow \\
\text{N} \\
\quad \text{fires} \\
\end{array}
\]

$p(t_1) = 4.2 \times 10^{-5}$ (preferred)

$\text{NP} \rightarrow \text{Det} \quad N \\
\text{S} \\
\downarrow \\
\text{NP} \rightarrow \text{VP} \\
\text{V} \rightarrow \text{fire} \\
\text{V} \rightarrow \text{ROOT s} \quad 0.086$

\[ t_1: \]

\[
\begin{array}{c}
\text{S} \\
\downarrow \\
\text{NP} \\
\downarrow \\
\text{Det} \\
\quad \text{the} \\
\downarrow \\
\text{N} \\
\quad \text{warehouse} \\
\downarrow \\
\text{V} \\
\quad \text{fires} \\
\end{array}
\]

$p(t_1) = 1.1 \times 10^{-5}$ (dispreferred)
Beam Search and Garden Path

• Prune low probability parses via beam search

• Assumption: if the relative probability of a parse with respect to the best parse drops below a certain threshold, it will be pruned

• Pruned parses are predicted to reflect garden-path sentences
Frame and Construction Probs

The horse raced past the barn fell.

\[ p(\text{race}, \langle \text{NP} \rangle) = 0.92 \]

\[ t_1: \]

\[
\begin{array}{c}
S \\
\text{NP} & \text{VP} \\
\text{the horse} & \text{raced}
\end{array}
\]

\[ p(t_1) = 0.92 \text{ (preferred)} \]

\[ p(\text{race}, \langle \text{NP NP} \rangle) = 0.08 \]

\[ \text{NP} \rightarrow \text{NP XP} \ 0.14 \]

\[ t_2: \]

\[
\begin{array}{c}
S \\
\text{NP} \\
\text{the horse} & \text{raced}
\end{array}
\]

\[ p(t_1) = 0.0112 \text{ (dispreferred)} \]
The bird found in the room died.

\[ p(\text{find, } \langle \text{NP} \rangle) = 0.38 \]

\[ t_1: \]
\[ S \]
\[ \text{NP} \quad \text{VP} \]
\[ \text{the bird} \quad \text{found} \]

\[ p(t_1) = 0.38 \text{ (preferred)} \]

\[ p(\text{find, } \langle \text{NP} \quad \text{NP} \rangle) = 0.62 \]

\[ \text{NP} \rightarrow \text{NP} \quad \text{XP} \quad 0.14 \]

\[ t_2: \]
\[ S \]
\[ \text{NP} \quad \ldots \]
\[ \text{NP} \quad \text{VP} \]
\[ \text{the bird} \quad \text{found} \]

\[ p(t_1) = 0.0868 \text{ (dispreferred)} \]
Claim: a tree is pruned, and therefore a garden-path, if the probability ratio is greater than 5:1.
Open Issues

• **Incrementality**: can we make more fine grained predictions about the time course of ambiguity

• **Relative difficulty**: Jurafsky doesn’t distinguish the relative difficulty of parses/interpretations that remain in the beam

• **Memory**: no account for memory load within a sentence (e.g. centre embeddings)

• **Coverage**: small, manually designed lexicon and grammar; tested on a handful of examples
A wide-coverage model: ICMM

- **ICMM**: Incremental Cascaded Markov Model (Crocker & Brants, 2000)
- Standard HMM POS tagger for lexical categories, similar to SLCM
- Structural probabilities computed as in a SCFG
- Wide coverage:
  - A fully implemented parser, trained on parsed corpora (Brown, WSJ, NEGRA)
  - Adapted to operate incrementally
• **Markov Models** for part-of-speech tagging use ‘horizontal’ probabilities (e.g., SLCM)

• **Stochastic CFGs** use ‘vertical’ probabilities (e.g., Jurafsky)

• **Cascaded Markov Models** apply ‘horizontal’ probabilities to levels higher than parts-of-speech
Incremental Cascaded Markov Models

• A parse consists of different layers of nodes

• Each Markov model layer consists of a series of nodes corresponding to phrasal (syntactic) categories

• Transitions correspond to trigram category probabilities

• Incremental (word-by-word) processing

• Build hypotheses for all layers as soon as a word is read

• Use each Markov model layer as a probabilistic filter, where only highest probability sequences are passed to the next layer
The warehouse prices
The warehouse prices the
The prices in the warehouse are

The prices of the beer are

The prices of the damage are

The prices of the beer are

The prices of the damage are

The prices of the damage are

The prices of the damage are

The prices of the damage are

The prices of the damage are

The prices of the damage are
The warehouse prices the beer
The prices the beer
The prices the beer warehousing cheaper.
The prices the beer warehouse cheaper
The prices the beer warehousing are cheaper (than the rest ...).
ICMM: Summary

- Advantages:
  - Wide coverage: accounts for a range of experimental findings concerning lexical and syntactic ambiguities
  - Cognitive plausibility: the model is incremental and uses limited memory

- Limitations:
  - Makes predictions about time course, but only at a coarse-grained level
  - Does not include verb subcategorization preferences
Summary & Conclusions

- **Motivation**: People process language: rapidly, robustly, and accurately
- Experimental evidence for probabilistic mechanisms
- **SLCM**: Simple, robust account of lexical category disambiguation
- **Jurafsky**: Probabilistic parser that models a range of local ambiguities
- **ICMM**: Incremental, broad coverage parser, combines SLCM & Jurafsky
Remaining Problems

• Integrating plausible parsing mechanisms:
  • Either bounded parallel, or serial (momentary parallel) with reanalysis

• Investigating more plausible `optimal functions’
  • More linguistically informed probabilistic models (lexical, semantic ...)
  • Integration with non-probabilistic decision strategies (e.g., recency)
  • More sophisticated integration of memory load constraints