Computational Psycholinguistics
Lecture 2: Probabilistic grammars and human sentence comprehension as incremental probabilistic parsing

Roger Levy
University of California – San Diego

ESSLLI
21 July 2009
Today

- Crash course in probability theory
- Crash course in natural language syntax and parsing
- Crash course in psycholinguistic methods
- Pruning models: Jurafsky 1996
Probability theory: what? why?

- Probability theory is the calculus of *reasoning under uncertainty*
- This makes it well-suited to modeling the process of language comprehension
- Language comprehension involves uncertainty about:
  - What has *already been said*

*The girl saw the boy with the telescope.*

- What has *not yet been said* (who has the telescope?)

*The children went outside to...* (play? chat? …)
Crash course in probability theory

- Event space $\Omega$
- A function $P$ from subsets of $\Omega$ to real numbers such that:
  - Non-negativity: $P(A) \geq 0, \forall A \subseteq \Omega$
  - Properness: $P(\Omega) = 1$
  - Disjoint union: $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
  - An improper function $P$ for which $P(\Omega) < 1$ is called deficient
Probability: an example

- Rolling a die has event space \( \Omega = \{1,2,3,4,5,6\} \)
- If it is a fair die, we require of the function \( P \):
  \[
  P(e) = \frac{1}{6}, \forall e \in \Omega
  \]
- Disjoint union means that this requirement completely specifies the probability distribution \( P \)
- For example, the event that a roll of the die comes out even is \( E = \{2,4,6\} \). For a fair die, its probability is
  \[
  P(E) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}
  \]
- Using disjoint union to calculate event probabilities is known as the counting method
Joint and conditional probability

- $P(X,Y)$ is called a *joint* probability
  - e.g., probability of a pair of dice coming out <4,6>
  - Two events are *independent* if the probability of the joint event is the product of the individual event probabilities:
    $$P(X, Y) = P(X)P(Y)$$
- $P(Y|X)$ is called a *conditional* probability
  - By definition, $P(X, Y) = P(Y|X)P(X)$
    $$= P(X|Y)P(Y)$$
  - This gives rise to Bayes’ Theorem:
    $$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$
Estimating probabilistic models

• With a fair die, we can calculate event probabilities using the counting method
• But usually, we can’t deduce the probabilities of the subevents involved
• Instead, we have to estimate them (=statistics!)
• Usually, this involves assuming a probabilistic model with some free parameters,* and choosing the values of the free parameters to match empirically obtained data

*(these are parametric estimation methods)
Maximum likelihood

- Simpler example: a coin flip
  - fair? unfair?
- Take a dataset of 20 coin flips, 12 heads and 8 tails
- Estimate the probability $p$ that the next result is heads
- Method of maximum likelihood: choose parameter values (i.e., $p$) that maximize the likelihood* of the data

$$L(X; p) = \binom{n}{k} p^k (1 - p)^{n-k}$$

- Here, maximum-likelihood estimate (MLE) is the relative-frequency estimate (RFE) $p = k/n$

*likelihood: the data’s probability, viewed as a function of your free parameters
Issues in model estimation

- Maximum-likelihood estimation has several problems:
  - Can’t incorporate a belief that coin is “likely” to be fair
  - MLEs can be *biased*
    - Try to estimate the number of words in a language from a finite sample
    - MLEs will always underestimate the number of words
  - There are other estimation techniques (Bayesian, maximum-entropy, …) that have different advantages
  - When we have “lots” of data,* the choice of estimation technique rarely makes much difference

*unfortunately, we rarely have “lots” of data
Generative vs. Discriminative Models

- Inference makes use of conditional probability distr’s

\[ P(H \mid O) \]

- Discriminatively-learned models estimate this conditional distribution directly

- Generatively-learned models estimate the joint probability of data and observation \( P(O, H) \)
  - *Bayes’ theorem* is used to find c.p.d. and do inference

\[
P(H \mid O) = \frac{P(O \mid H)P(H)}{P(O)}
\]
Generative vs. Discriminative in Psycholinguistics

- Different researchers have also placed the locus of action at generative (joint) versus discriminative (conditional) models.
- Are we interested in $P(\text{Tree}|\text{String})$ or $P(\text{Tree},\text{String})$?
- This reflects a difference in ambiguity type:
  - Uncertainty only about what *has been said*
  - Uncertainty also about what *may yet be said*
Today

• Crash course in probability theory
• Crash course in natural language syntax and parsing
• Crash course in psycholinguistic methods
• Pruning models: Jurafsky 1996
Crash course in grammars and parsing

• A *grammar* is a structured set of production rules
• Most commonly used for syntactic description, but also useful for (semantics, phonology, …)
• E.g., context-free grammars:

\[
\begin{align*}
S & \rightarrow \text{NP} \quad \text{VP} \\
\text{NP} & \rightarrow \text{Det} \quad \text{N} \\
\text{VP} & \rightarrow \text{V} \quad \text{NP}
\end{align*}
\]
\[
\begin{align*}
\text{Det} & \rightarrow \text{the} \\
\text{N} & \rightarrow \text{dog} \\
\text{N} & \rightarrow \text{cat} \\
\text{V} & \rightarrow \text{chased}
\end{align*}
\]
• A grammar is said to *license* a derivation

\[
\begin{align*}
\text{Det N} & \rightarrow \text{the dog} \\
\text{the cat}
\end{align*}
\]

\[
\begin{align*}
\text{Det N} & \rightarrow \text{the cat}
\end{align*}
\]
Top-down parsing

• Fundamental operation:
  
  \[
  S \rightarrow NP \ VP \\
  NP \rightarrow Det \ N \\
  Det \rightarrow The
  \]

• We saw this yesterday with Yngve (1960)
• Permits structure building inconsistent with perceived input, or corresponding to as-yet-unseen input

The coach smiled at the player tossed the frisbee.
Bottom-up parsing

- Fundamental operation: check whether a sequence of categories matches a rule’s *right-hand* side

  \[
  \begin{align*}
  \text{VP} & \rightarrow \text{V} \quad \text{NP} \\
  \text{PP} & \rightarrow \text{P} \quad \text{NP}
  \end{align*}
  \]

  \[
  \begin{align*}
  \text{S} & \rightarrow \text{NP} \quad \text{VP} \\
  \end{align*}
  \]

- Permits structure building inconsistent with global context

```
S
   /\  \
  /   \
VP   NP
   /\  \
  /   \
NP   V
    /\  \
   /   \
  NP   NP
      /\  \
     /   \
   the coach smiled at the player tossed the frisbee
```

Ambiguity

• There is usually more than one structural analysis for a (partial) sentence

  The girl saw the boy with...

• Corresponds to choices (non-determinism) in parsing
• VP can expand to V NP PP...
• …or VP can expand to V NP and then NP can expand to NP PP
• Ambiguity can be local (eventually resolved)…
  • …with a puppy on his lap.
• …or it can be global (unresolved):
  • …with binoculars.
Serial vs. Parallel processing

- A *serial* processing model is one where, when faced with a choice, chooses one alternative and discards the rest.
- A *parallel* model is one where at least two alternatives are chosen and maintained.
  - A *full parallel* model is one where *all* alternatives are maintained.
  - A *limited parallel* model is one where *some but not necessarily all* alternatives are maintained.

A joke about the man with an umbrella that I heard…

*ambiguity goes as the Catalan numbers (Church and Patel 1982)*
Dynamic programming

- There is an exponential number of parse trees for a given sentence (Church & Patil 1982)
- So sentence comprehension can’t entail an exhaustive enumeration of possible structural representations
- But parsing can be made tractable by *dynamic programming*
Dynamic programming (2)

• Dynamic programming = storage of partial results
• There are two ways to make an NP out of...

...but the resulting NP can be stored just once in the parsing process
• Result: parsing time polynomial (cubic for CFGs) in sentence length
• Still problematic for modeling human sentence proc.
Hybrid bottom-up and top-down

• Many methods used in practice are combinations of top-down and bottom-up regimens
• *Left-corner* parsing: incremental bottom-up parsing with top-down filtering
• *Earley* parsing: strictly incremental top-down parsing with top-down filtering and dynamic programming*

*solves problems of left-recursion that occur in top-down parsing*
A (generative) *probabilistic* grammar is one that associates probabilities with rule productions.

- e.g., a probabilistic context-free grammar (PCFG) has rule productions with probabilities like

  \[
  P(\text{NP} \rightarrow \text{Det N}) = 0.4 \\
  P(\text{NP} \rightarrow \text{NP PP}) = 0.23 \\
  P(\text{NP} \rightarrow \text{NP RC}) = 0.15 \\
  P(\text{NP} \rightarrow \text{NP and NP}) = 0.1 \\
  \ldots
  \]

- Interpret \(P(\text{NP} \rightarrow \text{Det N})\) as \(P(\text{Det N} \mid \text{NP})\)

- Among other things, PCFGs can be used to achieve *disambiguation* among parse structures
a man arrived yesterday

Total probability: $0.7 \times 0.35 \times 0.15 \times 0.3 \times 0.03 \times 0.02 \times 0.4 \times 0.07 = 1.85 \times 10^{-7}$
Probabilistic grammars (2)

- A derivation having zero probability corresponds to it being *unlicensed* in a non-probabilistic setting.
- But “canonical” or “frequent” structures can be distinguished from “marginal” or “rare” structures via the derivation rule probabilities.
- From a computational perspective, this allows probabilistic grammars to increase *coverage* (number + type of rules) while maintaining *ambiguity management*.
The probabilistic serial↔parallel gradient

- Suppose two incremental interpretations $I_{1,2}$ have probabilities $p_1 > 0.5 > p_2$ after seeing the last word $w_i$.
- A full-serial model might keep $I_1$ at activation level 1 and discard $I_2$ (i.e., activation level 0).
- A full-parallel model would keep both $I_1$ and $I_2$ at probabilities $p_1$ and $p_2$ respectively.
- An intermediate model would keep $I_1$ at $a_1 > p_1$ and $I_2$ at $a_2 < p_2$.
- (A “hyper-parallel” model might keep $I_1$ at $0.5 < a_1 < p_1$ and $I_2$ at $0.5 > a_2 > p_2$).
Today

- Crash course in probability theory
- Crash course in natural language syntax and parsing
- Crash course in psycholinguistic methods
- Pruning models: Jurafsky 1996
Psycholinguistic methodology

• The workhorses of psycholinguistic experimentation involve *behavioral* measures
  • What choices do people make in various types of language-producing and language-comprehending situations?
  • and how long do they take to make these choices?
• *Offline* measures
  • rating sentences, completing sentences, …
• *Online* measures
  • tracking people’s eye movements, having people read words aloud, reading under (implicit) time pressure…
Psycholinguistic methodology (2)

• [self-paced reading experiment demo now]
Caveat: *neurolinguistic* experimentation more and more widely used to study language comprehension

- methods vary in temporal and spatial resolution
- people are more passive in these experiments: sit back and listen to/read a sentence, word by word
- strictly speaking *not* behavioral measures
- the question of "what is difficult" becomes a little less straightforward
Today

- Crash course in probability theory
- Crash course in natural language syntax and parsing
- Crash course in psycholinguistic methods
- Pruning models: Jurafsky 1996
Pruning approaches

- Jurafsky 1996: a probabilistic approach to lexical access and syntactic disambiguation
- Main argument: sentence comprehension is probabilistic, construction-based, and parallel
- Probabilistic parsing model explains
  - human disambiguation preferences
  - garden-path sentences
- The probabilistic parsing model has two components:
  - *constituent* probabilities – a probabilistic CFG model
  - *valence* probabilities
• Every word is immediately completely integrated into the parse of the sentence (i.e., full incrementality)
• Alternative parses are ranked in a probabilistic model
• Parsing is limited-parallel: when an alternative parse has unacceptably low probability, it is pruned
• “Unacceptably low” is determined by beam search (described a few slides later)
Jurafsky 1996: valency model

- Whereas the constituency model makes use of only phrasal, not lexical information, the valency model tracks lexical subcategorization, e.g.:
  \[
P(\langle NP \ PP\rangle \mid discuss) = 0.24 \\
P(\langle NP\rangle \mid discuss) = 0.76
\]
  (in today’s NLP, these are called monolexical probabilities)

- In some cases, Jurafsky bins across categories:
  \[
P(\langle NP \ XP[+pred]\rangle \mid keep) = 0.81 \\
P(\langle NP\rangle \mid keep) = 0.19
\]
  where XP[+pred] can vary across AdjP, VP, PP, Particle…

*valence probs are RFES from Connine et al. (1984) and Penn Treebank*
The syntactic component of Jurafsky’s model is just probabilistic context-free grammars (PCFGs).

Total probability: $0.7 \times 0.35 \times 0.15 \times 0.3 \times 0.03 \times 0.02 \times 0.4 \times 0.07 = 1.85 \times 10^{-7}$
Ford et al. 1982 found effect of lexical selection in PP attachment preferences (offline, forced-choice):

- The women *discussed* the dogs on the beach
  - NP-attachment (the dogs that were on the beach) -- 90%
  - VP-attachment (discussed while on the beach) – 10%
- The women *kept* the dogs on the beach
  - NP-attachment – 5%
  - VP-attachment – 95%

Broadly confirmed in online attachment study by Taraban and McClelland 1988
• Jurafsky ranks parses as the *product* of constituent and valence probabilities:

\[
\begin{align*}
&[\text{root keep} \\
&\text{valence} \quad \langle \text{NP, XP [pred +]} \rangle \quad .81]
\end{align*}
\]

\[
\begin{array}{c}
\text{VP} \\
\text{V} \\
\text{NP} \\
\text{PP}
\end{array}
\]

\[
\begin{array}{c}
\text{keep} \\
\text{the dogs} \\
\text{on the beach}
\end{array}
\]

\[
(a) \quad .15 \times .81 = .12 \quad \text{(preferred)}
\]

\[
\begin{align*}
&[\text{root keep} \\
&\text{valence} \quad \langle \text{NP} \rangle \quad .19]
\end{align*}
\]

\[
\begin{array}{c}
\text{VP} \\
\text{V} \\
\text{NP}
\end{array}
\]

\[
\begin{array}{c}
\text{NP} \\
\text{PP}
\end{array}
\]

\[
\begin{array}{c}
\text{keep} \\
\text{the dogs} \\
\text{on the beach}
\end{array}
\]

\[
(b) \quad .19 \times .39 \times .14 = .01 \quad \text{(dispreferred)}
\]
Modeling offline preferences (3)

\[
\begin{align*}
\text{[root discuss valence } & <\text{NP,PP}> \ 0.24 \]\n\end{align*}
\]

\[
\begin{align*}
\text{VP} & \quad [0.15] \text{VP} \rightarrow V \ NP \ XP \\
\text{V} & \quad \text{discuss} \\
\text{NP} & \quad \text{the dogs} \\
\text{PP} & \quad \text{on the beach} \\
\end{align*}
\]

(a) \quad 0.15 * 0.24 = 0.036 \ (\text{dispreferred})

\[
\begin{align*}
\text{VP} & \quad [0.39] \text{VP} \rightarrow V \ NP \\
\text{V} & \quad \text{discuss} \\
\text{NP} & \quad \text{the dogs} \\
\text{PP} & \quad \text{on the beach} \\
\end{align*}
\]

\[
\begin{align*}
\text{[root discuss valence } & <\text{NP}> \ 0.76 \]\n\end{align*}
\]

(b) \quad 0.76 * 0.39 * 0.14 = 0.041 \ (\text{preferred})
Result

- Ranking with respect to parse probability matches offline preferences
- Note that only monotonicity, not degree of preference is matched
• Does this sentence make sense?
  *The complex houses married and single students and their families.*

• How about this one?
  *The warehouse fires a dozen employees each year.*

• And this one?
  *The warehouse fires destroyed all the buildings.*

• *Fires* can be either a noun or a verb. So can *houses*:
  \[
  [\text{NP The complex}] [\text{VP houses married and single students...}].
  \]

• These are *garden path* sentences

• Originally taken as some of the strongest evidence for *serial* processing by the human parser

*Frazier and Rayner 1987*
Limited parallel parsing

• Full-serial: keep only one incremental interpretation
• Full-parallel: keep all incremental interpretations
• Limited parallel: keep some but not all interpretations
• In a limited parallel model, garden-path effects can arise from the discarding of a needed interpretation

\[ S \ [NP \ The \ complex] \ [VP \ houses...] \ldots \]  \textcolor{green}{\leftarrow \text{discarded}}

\[ S \ [NP \ The \ complex \ houses \ ...] \ldots \] \textcolor{magenta}{\leftarrow \text{kept}}
Modeling online parsing: garden paths

- *Pruning* strategy for limited ranked-parallel processing
  - Each incremental analysis is ranked
  - Analyses falling below a threshold are discarded
  - In this framework, a model must characterize
    - The incremental analyses
    - The threshold for pruning
- Jurafsky 1996: partial context-free parses as analyses
- *Probability ratio* as pruning threshold
  - Ratio defined as $P(I) : P(I_{best})$
  - (Gibson 1991: *complexity ratio* for pruning threshold)
Garden path models 1: N/V ambiguity

- Each analysis is a partial PCFG tree
- *Tree prefix probability* used for ranking of analysis

```
  S
 /\  \\
NP VP
 /|
DT NN V
```

- These nodes are actually still undergoing expansion

- Partial rule probs *marginalize* over rule completions

\[
P(VP \rightarrow V \ldots) = \sum_{\alpha} P(VP \rightarrow V \alpha)
\]
N/V ambiguity (2)

- Partial CF tree analysis of the complex houses...

(a) (preferred) $1.2 \times 10^{-7}$

(b) (dispreferred) $4.5 \times 10^{-10}$

- Analysis of houses as noun has much lower probability than analysis as verb (> 250:1)
- Hypothesis: the low-ranking alternative is discarded
N/V ambiguity (3)

• Note that top-down vs. bottom-up questions are immediately implicated, in theory
• Jurafsky includes the cost of generating the initial NP under the S
  • of course, it’s a small cost as \( P(S \rightarrow NP \ldots) = 0.92 \)
• If parsing were bottom-up, that cost would not have been explicitly calculated yet
Garden path models 2

- The most famous garden-paths: reduced relative clauses (RRCs) versus main clauses (MCs)

The horse raced past the barn fell.

(that was)

- From the valence + simple-constituency perspective, MC and RRC analyses differ in two places:

transitive valence: $p=0.08$

best intransitive: $p=0.92$
• 82 : 1 probability ratio means that lower-probability analysis is discarded

• In contrast, some RRCs do not induce garden paths:
  
  *The bird found in the room died.*

• Here, the probability ratio turns out to be much closer ($\approx 4 : 1$) because *found* is preferentially transitive

• Conclusion within pruning theory: *beam threshold is between 4 : 1 and 82 : 1*

• (granularity issue: when exactly does probability cost of valence get paid???)
Notes on the probabilistic model

Jurafsky 1996 is a *product-of-experts* (PoE) model

\[ P(X) = \frac{1}{Z} \prod_i P_i(X) \]

- Expert 1: the constituency model
- Expert 2: the valence model

PoEs are flexible and easy to define, but hard to learn

- The Jurafsky 1996 model is actually *deficient* (loses probability mass), due to relative frequency estimation

\[
\sum_i P(valence_i|discuss) = P(NP\ PP|discuss)P(VP \rightarrow V \ NP \ XP) \\
\quad + P(NP|discuss)P(VP \rightarrow V \ NP) \\
\quad = 0.15 \times 0.24 \\
\quad + 0.76 \times 0.39 \\
\quad = 0.036 + 0.2964 \leq 1
\]
Notes on the probabilistic model (2)

- Jurafsky 1996 predated most work on lexicalized parsers (Collins 1999, Charniak 1997)
- In a generative lexicalized parser, valence and constituency are often combined through decomposition & Markov assumptions, e.g.,

\[ P(\text{valence}, \text{verb}|VP) = P(\text{head} = \text{verb}|VP)P(\text{valence}|VP, \text{verb}) \]

- The use of decomposition makes it easy to learn non-deficient models
Jurafsky 1996 & pruning: main points

- Syntactic comprehension is probabilistic
- Offline preferences explained by syntactic + valence probabilities
- Online garden-path results explained by same model, when beam search/pruning is assumed
General issues

- What is the granularity of incremental analysis?
  - In $[_{NP} \textit{the complex houses}]$, \textit{complex} could be an adjective (=\textit{the houses are complex})
  - \textit{complex} could also be a noun (=\textit{the houses of the complex})
  - Should these be distinguished, or combined?
  - When does valence probability cost get paid?

- What is the criterion for abandoning an analysis?

- Should the \textit{number} of maintained analyses affect processing difficulty as well?
Main-verb/reduced-relative (MV/RR) ambiguity again

Variant of the famous garden-path sentence

- The cop arrested by the detective was guilty
- The cop that was arrested by the detective was guilty
- The crook arrested by the detective was guilty
- The crook that was arrested by the detective was guilty

Ambiguity at the first verb is resolved at the PP

But the viability of RR versus MC interpretations at the temporary ambiguity is affected by thematic fit
Garden-path gradience

- Empirical data, McRae et al., 1998

The cop arrested... was guilty

RT pattern here is the crucial result (why?)
• Narayanan & Jurafsky 1998 was a *pruning* model: predicted increased RT when the required incremental parse had been discarded

• Narayanan & Jurafsky 2002 adds the idea of *attention shift*: predicts increased RT when the highest-ranked analysis changes status

• Their answers to conceptual issues:
  • Granularity: only RR and MV analyses are maintained
  • Ranking metric: probability theory (*some novelty here*)
  • Parallelism: full (but only two analyses!)
# Ranking analyses: building blocks

<table>
<thead>
<tr>
<th>Data</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(\text{Agent} \mid \text{verb, initial NP})$</td>
<td>McRae et al. (1998)</td>
</tr>
<tr>
<td>$P(\text{Patient} \mid \text{verb, initial NP})$</td>
<td>McRae et al. (1998)</td>
</tr>
<tr>
<td>$P(\text{Participle} \mid \text{verb})$</td>
<td>British National Corpus counts</td>
</tr>
<tr>
<td>$P(\text{SimplePast} \mid \text{verb})$</td>
<td>British National Corpus counts</td>
</tr>
<tr>
<td>$P(\text{transitive} \mid \text{verb})$</td>
<td>TASA corpus counts</td>
</tr>
<tr>
<td>$P(\text{intransitive} \mid \text{verb})$</td>
<td>TASA corpus counts</td>
</tr>
<tr>
<td>$P(\text{RR} \mid \text{initial NP, verb-ed, by})$</td>
<td>McRae et al. (1998) (.8, .2)</td>
</tr>
<tr>
<td>$P(\text{RR} \mid \text{initial NP, verb-ed, by, the})$</td>
<td>McRae et al. (1998) (.875, .125)</td>
</tr>
<tr>
<td>$P(\text{Agent} \mid \text{initial NP, verb-ed, by, the, NP})$</td>
<td>McRae et al. (1998) (4.6 average)</td>
</tr>
<tr>
<td>$P(\text{MC} \mid \text{SCFG prefix})$</td>
<td>SCFG counts from Penn Treebank</td>
</tr>
<tr>
<td>$P(\text{RR} \mid \text{SCFG prefix})$</td>
<td>SCFG counts from Penn Treebank</td>
</tr>
</tbody>
</table>
Ranking analyses via Bayes nets

\[
V = \text{examine-ed} \quad \text{type_of(Subj)} = \text{witness}
\]

\[
P(A|v\), \quad P(T|v)
\]

\[
\text{Arg} \quad \text{Tense} \quad \text{Semantic_fit}
\]

\[
M_{\text{thm}} \quad Tense = \text{past} \quad \text{Sem_fit} = \text{Agent}
\]

\[
R_{\text{thm}} \quad \text{Arg} = \text{trans} \quad Tense = \text{pp} \quad \text{Sem_fit} = \text{Theme}
\]
Noisy AND-gates

- A technique for constraining the form of a conjunctive probability distribution $P(X|U_1,\ldots,U_i,\ldots,U_n)$

- **Accountability**: If all $U_i$ are true, then $X$ is true

- **Exception independence**: Each $U_i$ that is false has a chance $1-q_i$ of making $X$ false

- If we call $F_u$ the set of $U_i$ that are false, then

  \[
  P(X = \text{True}|F_u) = \prod_{u \in F_u} q_i
  \]

  \[
  P(X = \text{False}|F_u) = 1 - \prod_{u \in F_u} q_i
  \]

- This facilitates learning as now there are only $n$ parameters, instead of $2^n$

*Pearl 1988*
Modeling results

- Posterior probabilities “flip” only in good-agent case

![Graph showing the comparison between cop arrested by the FLIPS and crook arrested by the FLIPS. The graph illustrates the change in posterior probabilities for different scenarios.]
• Goes beyond Jurafsky 1996 in two respects
  • use of Bayes nets to formalize the probabilistic relationships between different types of evidence
  • posits attention shift as well as pruning as a source of processing difficulty
For Wednesday

- Read Hale, 2001; and Levy, Reali, & Griffiths, 2009