Computational Psycholinguistics

Lecture 6: Probabilistic Models of Human Sentence Processing

Afra Alishahi

December 8, 2008

Symbolic Models

- So far: symbolic accounts of processing
  - Recover sentence structure incrementally
  - Use structure- or grammar-based strategies or additional information sources to resolve ambiguity
  - Predict garden-path when reanalysis is needed due to memory limitations or choosing the wrong strategy
- But,
  - Do not explain empirical findings about the role of linguistic experience

The Role of Experience

- Experimental findings:
  - Frequency information plays a central role in disambiguation
  - People can deal with complexity and ambiguity accurately and in real time, despite the limitation of their cognitive resources
- Alternative: probabilistic approaches
  - Develop experience-based models, which draw on previous exposure to language

Probabilistic Models

- Probabilistic methods are widely used in
  - decision making under uncertainty
  - dealing with noisy and ambiguous data
  - various areas of computational linguistics
- Probability theory was invented as a cognitive model of human reasoning under uncertainty
  - But is human language processing an optimal, rational process?

Probabilistic Comprehension

- Probabilistic human sentence processing: what does it mean?
  - Accessibility: more probable structures are accessed more quickly, or with less evidence
  - Disambiguation: more probable interpretations are more likely to be chosen
  - Processing difficulty: certain interpretations have particularly low probabilities, or sudden switches of probabilistic preference between interpretations

Evidential Reasoning

- Probabilistic modeling of comprehension:
  - A principled algorithm for weighting and combining evidence to choose interpretations in comprehension
- Bayes' rule: break down complex probabilities into ones that are easier to compute

$$P(i|e) = \frac{P(e|i)P(i)}{P(e)}$$

Probability of an interpretation i given some evidence e

Likelihood of evidence e assuming an interpretation i

Prior probability of interpretation i

Prior probability of evidence e

➡ Find the interpretation i that maximizes $P(i|e)$

Prior Knowledge: Frequency

- How to estimate the prior probability of an interpretation, or $P(i)$?
  - Relative frequency: more frequent structures have higher prior probability
  - Complex structures are too rare, so their probability cannot be estimated directly.
  - Independence assumptions: estimate the probability of a complex structure from the counts of smaller parts

➡ Probabilistic model predicts frequency effects for various kinds of structures.

Lexical Frequency Effects

- Word frequency:
  - More frequent words are accessed and articulated more quickly.
- Sense and category disambiguation:
  - Frequency of syntactic and semantic categories associated with words affect comprehension.
- Subcategorization frame selection:
  - Frequencies of verb subcategorization frames play a role in disambiguation.

Estimating Lexical Frequency

- Word frequency plays a robust effect in lexical comprehension and production
  - Frequencies are usually gathered from an annotated corpus, e.g. Brown corpus of American English
  - Problems: the corpus is old, provides production data not representative of the daily exposure to language
  - Yet strong frequency effects have been found
    - frequencies from different corpora are correlated
    - Broad-grained frequencies are used.

Joint & Conditional Probabilities

- Probability of a given word given its neighbors plays a role in comprehension and production
- Joint probability of two words:

$$P(w_{i-1}w_i) = \frac{C(w_{i-1}w_i)}{N}$$

- Conditional probability of a word given a previous word:

$$P(w_i|w_{i-1}) = \frac{C(w_{i-1}w_i)}{C(w_{i-1})}$$

Empirical Evidence

- MacDonald (1993): joint (word-pair) probabilities of word pairs affect reading times:

  The doctor refused to believe that the shrine cures people of many fatal diseases ...

  The doctor refused to believe that the miracle cures people of many fatal diseases...

- MacDonald (2001): conditional (bigram) probability is good predictor of gaze on a word.
- Bod (2000): frequent three-word (SVO) sentences are easier and faster to recognize.

Probabilistic Comprehension

- A rational approach to language processing:
  - Identify the goal of the process
  - Reason about the function that best achieves the goal
  - ➡ Choose parsing operations that maximize the likelihood of finding the intended interpretation

$$interpretation = \text{argmax}_i P(i|e) = \text{argmax}_i P(e|i)P(i)$$

- What evidence?

$$interpretation_j = \text{argmax}_i P(i|w_{1\ldots j}, K)$$

words seen so far in the sentence

Our general knowlege

Lexical Category Disambiguation

- Much of the ambiguity in syntactic processing derives from ambiguity at the lexical level.
- 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.

Probabilistic Lexical Processing

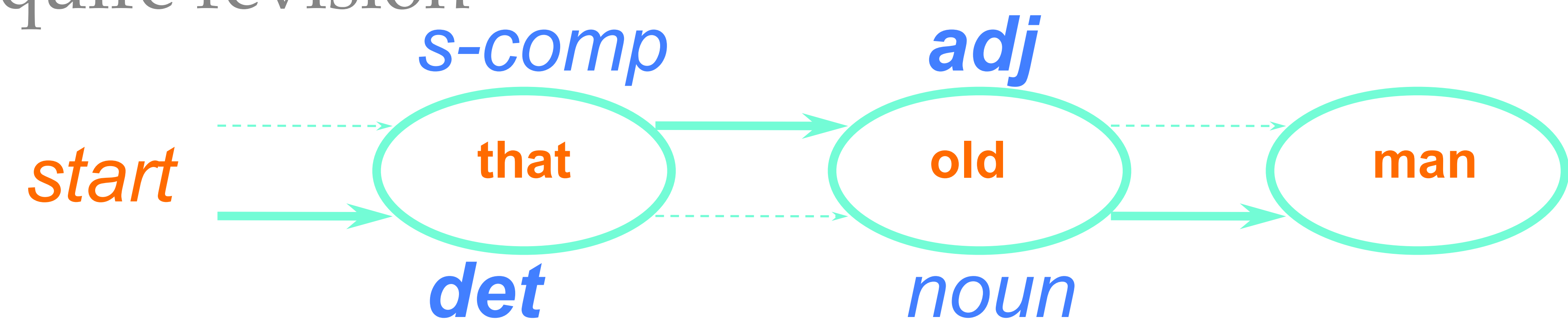
- SLCM (Corley & Crocker 2000): a simple POS tagger
  - Find the best category path for a sequence of words

$$P(t_0,\ldots t_n, w_0,\ldots w_n) \approx \prod_{i=1}^{n} P(w_i \mid t_i)\, P(t_i \mid t_{i-1})$$

lexical bias — $P(w_i \mid t_i)$

category context — $P(t_i \mid t_{i-1})$

  - Categories are assigned incrementally: Best path may require revision



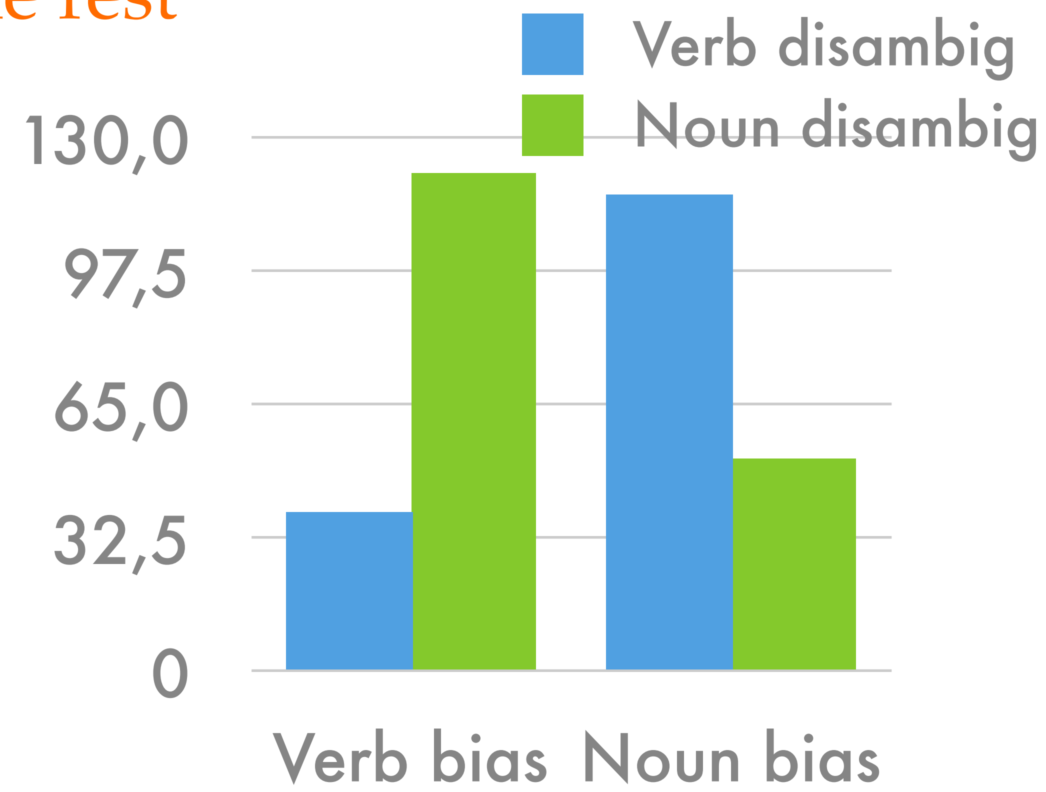
SLCM Predictions

- The Statistical Hypothesis:
  - Lexical word-category frequencies are used for initial category resolution
- The Modularity Hypothesis:
  - Initial category disambiguation is modular, and not determined by (e.g. syntactic) context

Statistical Lexical Categorization

the warehouse *prices* the beer very modestly
DET N N / V **V!**
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!** ...

➡ Lexical category frequency determines initial decisions



Modular Disambiguation

- *That* ambiguity

  The doctor told the woman that ...

  story.

  diet was unhealthy.

  he was in love with her husband.

  he was in love with to leave.

  story was about to leave.

- Which factors contribute to disambiguation?

Modular Disambiguation

- Juliano & Tanenhaus (1993): ambiguous words are resolved by their preceding context.

  That [DET] *experienced diplomat(s) would be very helpful ...*

  The lawyer insisted that [COMP] *the experienced diplomat(s) would be very helpful ...*

  Initially: DET =.35, COMP=.11
  Post-verbally: COMP=.93, DET =.06

  - Reading times increase when dispreferred interpretation (according to context) is forced.

Subcategorization Frequencies

The doctor remembered [NP *the idea*].

The doctor remembered [S *that the idea had already been proposed*].

The doctor suspected [NP *the idea*].

The doctor suspected [S *that the idea would turn out not to work*].

- Both verbs allow both subcategorization frames, but with different frequencies.
  - Frequencies are an estimate of the conditional probability of the frame given the verb: *P(frame|verb)*

Subcategorization Frequencies

- Conditional probabilities have been shown to play a role in disambiguation
  - Jurafsky (1996), Trueswell et al. (1993), MacDonald (1994)

The sleek greyhound raced at the track won four trophies.

The sleek greyhound admired at the track won four trophies.

  - Reduced relative clause constructions are easier to process when the word following the ambiguous verb matches the verb's transitivity bias.

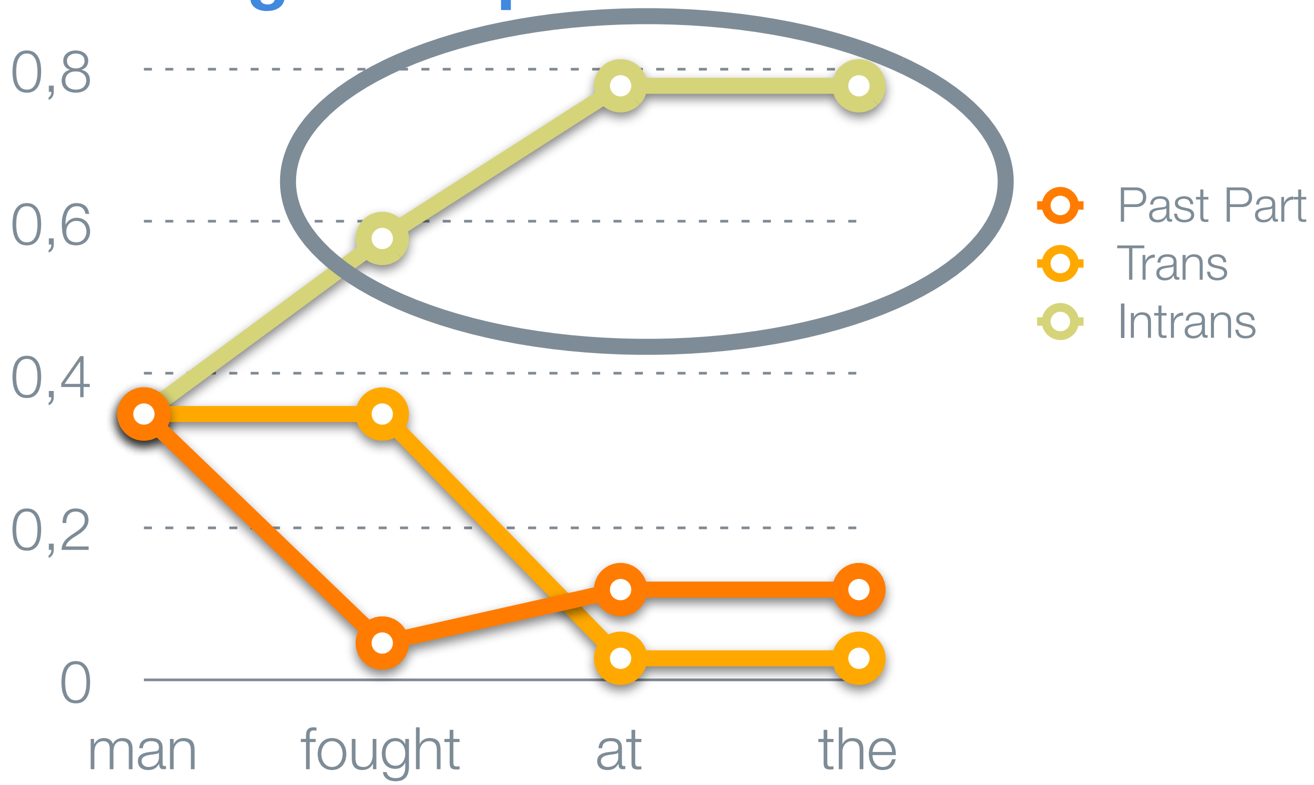
SCLM Account

- Using corpus-based verb frequencies:

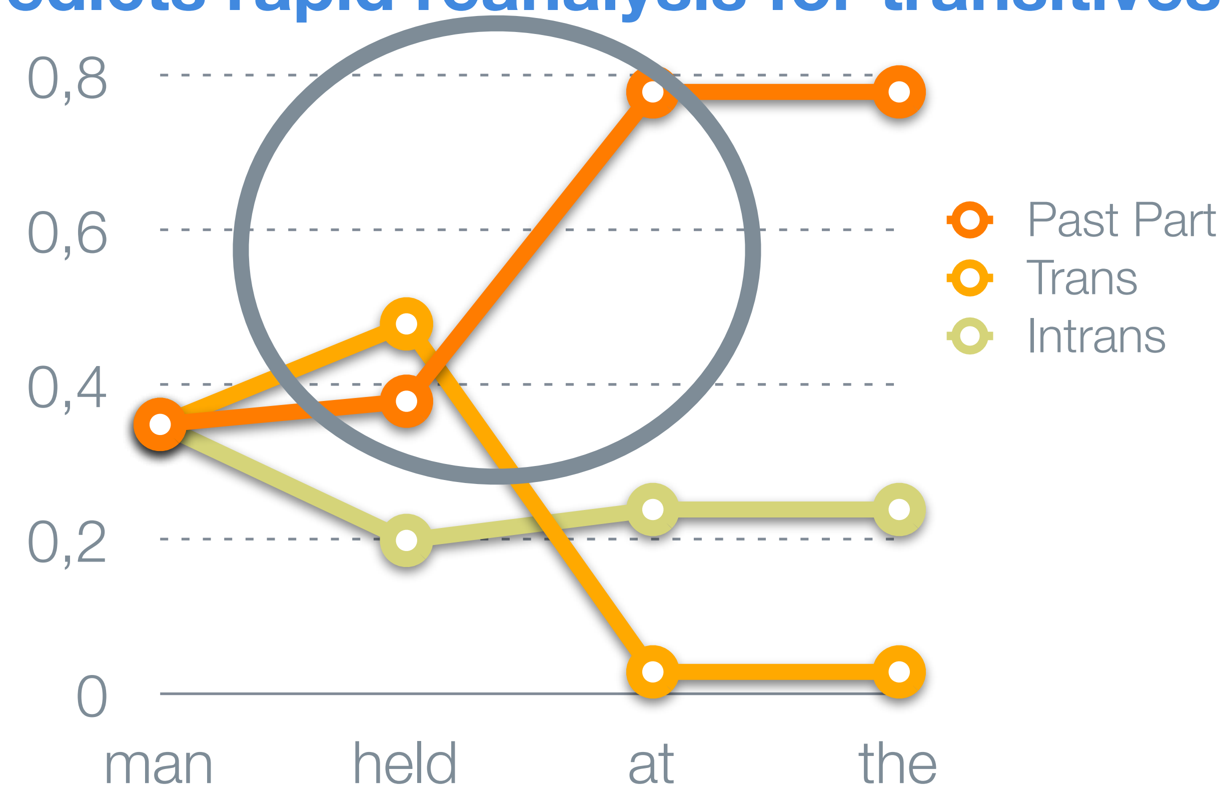
The man fought *at the police station fainted.* [intransitive]

The man held *at the police station fainted.* [transitive]

Predicts garden path for intransitives



Predicts rapid reanalysis for transitives



SLCM Summary

- Psychological plausibility:
  - relatively lower statistical complexity, high accuracy
- Clear predictions:
  - Statistical: frequency drives initial category decisions
  - Modular: syntax structure does not determine initial category decisions
  - Indication of which features are exploited (e.g. transitivity, but not number)

➡ Not a model of syntactic processing.

Probabilistic Syntax Processing

- Empirical findings
  - Statistical biases for non-lexical syntactic structures
- Which probabilities to look at?
  - The grain problem: the appropriate units to count
- How to use probabilities in sentence processing?
  - Associate grammatical knowledge with probabilistic weights
  - Statistical processing mechanisms: probabilistic parsing operations

Marr's Levels of Modeling

- Marr (1982) identifies three levels of describing cognitive processes:
  - Computational level: defines *what* is computed
  - Algorithmic level: specifies *how* computation takes place
  - Implementation level: states how the algorithms are actually *realized* in brain

Probabilistic Models of Language

- The rational models of language processing provide theories at the computational level
  - The likelihood function defines the goal of the process
  - Each model defines the algorithmic instantiation
    - It can be modified without changing the computational theory
    - E.g. bigram model and Viterbi algorithm in SLCM
  - The implementation level is usually not provided.