## Computational Psycholinguistics

## Lecture 8: Rational Analysis of Parsing

## Matthew W Crocker

Computerlinguistik<br>Universität des Saarlandes

## Probabilistic Models, so far ...

We have argued for probabilistic models because:

- Psychological evidence for frequency effects:
+ Word category \& sense, subcategorization, attachment (?)
$\square$ Rational: probabilistic techniques explain the fact that people process language rapidly, accurately, and robustly.
- Interesting for modular architectures, where statistics provide good "heuristics" in the absence of full knowledge.
$\square$ Three models, explain both good performance \& "pathologies"
- SLCM: a hidden Markov model of lexical category disambiguation
- Jurafsky: probabilistic models of parsing and lexical access + Combines structure \& frame probabilities, not "fully implemented".
- ICMM: implementation of a wide-coverage probabilistic parser:
+ Combines "phrase structure", and "phrase sequence" probabilities
Criticisms of :
- high performance probabilistic parsers are typically massively parallel and also nonincremental.
- practical concerns require the estimation of probabilities in ways which people may not need to, can we reason about "true likelihood" ?


## More on Probabilistic Models

Psychological plausibility of wide-coverage probabilistic parsers:
How do memory restrictions and strict incrementality affect performance?
Brants \& Crocker (2000)

Probabilistic implementations contain many practical simplifications concerning:
$\square$ How sub-probabilities are combined
How probabilities estimated are from corpora

Stepping Back: considering probabilistic accounts generally

- Serial probabilistic parsing
- Parallel, Bayesian parsing

Criticism: are likelihood models „optimal", do they account for the data
$\square$ An alternative rational analysis: Informativity Theory

## Psychological Plausibility

Are wide-coverage, probabilistic models cognitively plausible?

- Models: Jurafsky (1996); Crocker \& Brants (2000)

Cognitive constraints: Memory and Incrementality

- Broad coverage probabilistic parsers:
- High accuracy: 86\% precision/recall
- Robust: Analyse all and ill-formed input
- Non-incremental
- Massive parallelism
$\square$ ICMM is a broad coverage, probabilistic parser:
- Restricted beam

I Incremental processing
What is the general performance of probabilistic parser that:

- Has restricted memory resources

Strictly incremental parsing (and pruning)

## Design of the Experiment

Adapted a standard Stochastic Context Free Grammar:
Generality of results (not just for ICMM), not highest performance

- Incremental Processing + No look-ahead: full processing on each word
+ Immediate pruning: reduces memory requirements
+ Simple ranking strategy
- Pruning: active/inactive/both
+ Variable Beam: edges close to best are kept
+ Fixed Beam: fixed number of best edges are kept

Training: Wall street journal sections 2-21

- Testing: From section 22
$\square 1578$ sentences of length 40 or less


## Results for Incremental SCFG

Baseline performance:
$\left.\begin{array}{l}\text { Recall: } 68.82 \% \\ \text { Precision: } 73.77 \%\end{array}\right\}$ F-Score: 71.21

- Chart size: 141,650
- Avg \# of analysis per span: 18.7

Speed: 1.8 Tokens/Sec

Restricted model:
Recall: 68.82\% $\}$ F-Score: 71.16Precision: 73.66\%Chart size: 1.15\%Avg \# of analysis per span: 2Speed: 301 Tokens/Sec

- Fixed beam (inactive: 2 active: 4)


## A Simple Likelihood Account

■ Can we reason about the behaviour of a 'pure' likelihood model?

A simple probabilistic model:
At each point of ambiguity, simply select the structure with the greatest probability
Consider the sentence fragment:

- A) "The athlete realized his shoes were out of reach"

B B) "The athlete realized his goals yesterday"

- Priors: $\quad \mathrm{P}($ Hdo|realized $)=0.2 \quad \mathrm{P}($ Hes $\mid$ realized $)=0.8$
$\square$ When "his" is encountered, construct both the direct object and embedded sentence structures are built, and Hes is adopted.
- Prediction:
- (A) should be easy, (B) should require reanalysis after DO phrase.


## A Bayesian Model

An 'Ideal' probabilistic model:
Incrementally determine probabilities for all possible structures
At a point of ambiguity, rank structures according to prior probabilities
As new words are found, use posterior probabilities (Bayes' Theorem):

$$
p\left(H_{i} \mid E\right)=\frac{P\left(E \mid H_{i}\right) P\left(H_{i}\right)}{P\left(E \mid H_{i}\right) P\left(H_{i}\right)+P\left(E \mid H_{j}\right) P\left(H_{j}\right)}
$$

Consider the sentence fragment: "The athlete realized his ...

- Priors: $\quad \mathrm{P}(\mathrm{Hdo})=0.2 \quad \mathrm{P}(\mathrm{Hes})=0.8$
- Conditional Probability of $\mathrm{P}(\mathrm{E} \mid \mathrm{H})$

| $\mathrm{P}($ goals $\mid$ Hdo $)=0.2$ | $\mathrm{P}($ goals $\mid$ Hes $)=0.0001$ |
| :--- | :--- |
| $\mathrm{P}($ shoes $\mid$ Hdo $)=0.00001$ | P (shoes $\mid$ Hes $)=0.0001$ |
| $\mathrm{P}(\mathrm{X} \mid \mathrm{H} d o)=0.79999$ | $\mathrm{P}(\mathrm{X} \mid$ Hes $)=0.9998$ |

## Behaviour of Bayesian Model

When realized is encountered, hypothesise $\mathrm{H}_{\mathrm{es}}$ :

|  | $\mathrm{P}(\mathrm{H} d o)$ | $\mathrm{P}(\mathrm{Hes})$ |  |
| :---: | :---: | :---: | :---: |
| P (goals) | $\begin{gathered} \mathrm{P}(\text { goals } \mid \mathrm{H} d o) \mathrm{P}(\mathrm{H} d o) \\ .2 \times .2=.04 \end{gathered}$ | $\begin{gathered} \mathrm{P}(\text { goals } \mid \mathrm{Hes}) \mathrm{P}(\mathrm{Hes}) \\ .0001 \times .8=.00008 \end{gathered}$ | . 04008 |
| P (shoes) | $\begin{gathered} \mathrm{P}(\text { shoes } \mid \mathrm{H} d o) \mathrm{P}(\mathrm{H} d o) \\ .00001 \times .2=.000002 \end{gathered}$ | $\begin{gathered} \mathrm{P}(\text { shoes } \mid \mathrm{Hes}) \mathrm{P}(\mathrm{Hes}) \\ .0001 \times .8=.00008 \end{gathered}$ | . 000082 |
| P (Other) | $\begin{gathered} \mathrm{P}(\text { other } \mathrm{H} d o) \mathrm{P}(\mathrm{H} d o) \\ .79999 \times .2=.159998 \end{gathered}$ | $\begin{gathered} \mathrm{P}(\text { other } \mathrm{H} e s) \mathrm{P}(\mathrm{Hes}) \\ .9998 \times .8=.79984 \end{gathered}$ | . 959838 |
|  | . 2 | . 8 | 1.0 |

When new evidence is seen, compute $\mathrm{P}(\mathrm{H} \mid \mathrm{E})$ :

$$
p\left(H_{i} \mid E\right)=\frac{P\left(E \mid H_{i}\right) P\left(H_{i}\right)}{P\left(E \mid H_{i}\right) P\left(H_{i}\right)+P\left(E \mid H_{j}\right) P\left(H_{j}\right)}
$$

- If shoes then:

$$
\begin{array}{|lc|}
\hline \mathrm{P}\left(\mathrm{H}_{\text {es }} \text { shoes }\right)=.98 & \mathrm{P}\left(\mathrm{H}_{\text {do }} \text { |shoes }\right)=.02 \\
\hline \mathrm{P}\left(\mathrm{H}_{\text {es }} \text { |goals }\right)=.002 & \mathrm{P}\left(\mathrm{H}_{\text {do }} \text { goals }\right)=.998 \\
\hline
\end{array}
$$

Behaviour: Should get the globally preferred analysis ...
$\square$ Locally predicts initial preference for $\mathrm{H}_{\mathrm{es}}$.
Correctly "switches" to $H_{d o}$ based on new evidence.
Assumes full parallelism: psychologically implausible?

## Likelihood prediction for NP/S



## More Problems for Likelihood

NP/Z Complement Ambiguity: As the professor lectured the students ...


- Likelihood predictions:

When NP is encountered, build more likely (intransitive) structure + No difficulty if VP is then encountered (above) + Reanalysis effect only if second NP appears

- "As the professor lectured the students the sparrows became restless"

■ Experimental evidence: (Pickering, Traxler \& Crocker, 2000)
$\square$ Opposite, to above!

## Refining the Rational Analysis: Informativity

How can we explain the preference for object attachment (i.e. the NP/S and NP/Z findings) within a rational framework?

- Properties of the incremental parsing mechanism:
$\square$ local ambiguities $L_{i}$ must be resolved as they are encountered:
success $=$ settling on the globally correct analysis
$P($ global success $)=\prod_{i=1}^{n} P\left(\right.$ success at $\left.L_{i}\right)$
+ Initially adopting an analysis, which is ultimately correct
* Backing-out of a wrong analysis, and settling on the correct one
- Computational assumptions:
local reanalysis is often easy, long-distance reanalysis is difficult
$\square$ only one (or few) interpretations can be 'foregrounded'
■ Foreground the analysis which can be most confidently "tested".
Increase probability of locally backing out of a wrong analysis
Avoid being led down the garden path by pure likelihood


## Deriving the optimal function: Informativity

$\Rightarrow$ Informativity:
$\Rightarrow \mathrm{I}=\mathrm{f}(\mathrm{P}, \mathrm{T})$

- $\mathrm{P}=$ prior probability $\mathrm{T}=$ testability
- Ideally:
- Priors: are based on our experience
- Testability: measures how useful new evidence $E$ will be in estimating $P(H \mid E)$.
- $\mathrm{P}\left(\right.$ Pass $\left._{i}\right)=$ probability that evidence confirms $\mathrm{H}_{i}$
- We define Specificity for $\mathrm{H}_{i}$ as:
$\square S_{i}=1 / \mathrm{P}\left(\right.$ Pass $\left._{i}\right)$

Rational behaviour: maximise the chance of making the correct analysis, soon.

```
The Derivation:
Consider two hypotheses }\mp@subsup{\textrm{H}}{1}{}&\mp@subsup{\textrm{H}}{2}{}\mathrm{ :
    P(\mp@subsup{Correct }{1}{2})=P(\mp@subsup{H}{1}{},\mp@subsup{\mathrm{ Pass }}{1}{})+P(\mp@subsup{H}{2}{},\mp@subsup{\mathrm{ Fail }}{1}{})
    =P(\mp@subsup{Pass}{1}{}|\mp@subsup{H}{1}{})P(\mp@subsup{H}{1}{})+P(\mp@subsup{F}{\mathrm{ Fail }}{1}|}|\mp@subsup{H}{2}{})P(\mp@subsup{H}{2}{}
    =P(\mp@subsup{H}{1}{})+(1-1/\mp@subsup{S}{1}{})P(\mp@subsup{H}{2}{2}
    P(\mp@subsup{Correct }{2}{2})=P(\mp@subsup{\textrm{H}}{2}{})+(1-1/\mp@subsup{S}{2}{}})P(\mp@subsup{H}{1}{}
Choose }\mp@subsup{\textrm{H}}{\textrm{i}}{}\mathrm{ where P(Correcti) greatest:
    P(\mp@subsup{Correct }{1}{})>P(\mp@subsup{\mathrm{ Correct }}{2}{})
    P(H1)+(1-1/S ( )P(H2)>
        P(H2)+(1-1/S S)P(H1)
    S
So, choose H}\mp@subsup{H}{i}{}\mathrm{ where SiP(H
    maximised
```


## NP/S Revisited

Pickering, Traxler \& Crocker: NP vs. S
The athlete realised his shoes were out of reach The athlete realised his goals were out of reach For a set of S-bias verbs (corpus \& completion).

Eye-tracking study revealed:
$\rightarrow$ Increased RTs in coloured region

Consistent with initial object attachment
$\checkmark$ Confirms the prediction of the Informativity Model
$x$ Falsifies the analysis based on strict Maximum Likelihood.

## Estimating Informativity: An example

Choose $H_{i}$ where $S_{i} P\left(H_{i}\right)$ is maximised

$$
P\left(H_{i}\right)=\frac{f\left(H_{i}\right)}{\sum_{\forall j} f\left(H_{j}\right)} \quad S_{i}=\frac{\text { CorpusSize }}{\sum_{\forall w j \in \text { Passi }} f\left(w_{j}\right)}
$$

- Extract 100 tokens for each verb, from the BNC using GSEARCH
- Then estimate $P$ and $S$ as above:
- $\mathrm{P}\left(\mathrm{H}_{\mathrm{NP}} \mid\right.$ verb $)=0.3 \quad \mathrm{P}\left(\mathrm{H}_{\mathrm{s}} \mid\right.$ verb $)=0.55 \quad \mathrm{P}\left(\mathrm{H}_{\mathrm{X}} \mid\right.$ verb $)=0.15$
- Specificity: $S_{N P}$ is underestimated due to small corpus counts
- as |Corpus) increases, the number of words that Pass will not increase as quickly for $\mathrm{S}_{\mathrm{NP}}$ as for $\mathrm{S}_{\mathrm{S}}$

| Specificity: | Object | Subject |
| :--- | ---: | ---: |
| admit | 105 | 30 |
| decide | 90 | 399 |
| hint | 1187 | 363 |
| hmply | 352 | 30 |
| pretend | 896 | 122 |
| realise | 81 | 45 |
| Total | 2711 | 989 |

## Further Predictions

General preference for argument attachment over non-argument
$\square$ Since selectional restriction will correlate with Informativity
$\square$ Prefers formation of dependencies with existing structure:
$\square$ Clause boundary: "When John walks the fish jump"
$\square$ NP PP V: "The girl from Holland laughed/arrived" [Dutch]

When Specificity is constant, use Priors:

- "Tuning" effects in modifier attachment, E.g. Relative clauses

Possibly lexically specific cases:
$\square$ PP-attachment:

+ "I ate the pizza with pepperoni/a fork"
+ "I saw the man with the moustache/the binoculars"

