Lecture 1: Generative Models & Statistical Inference

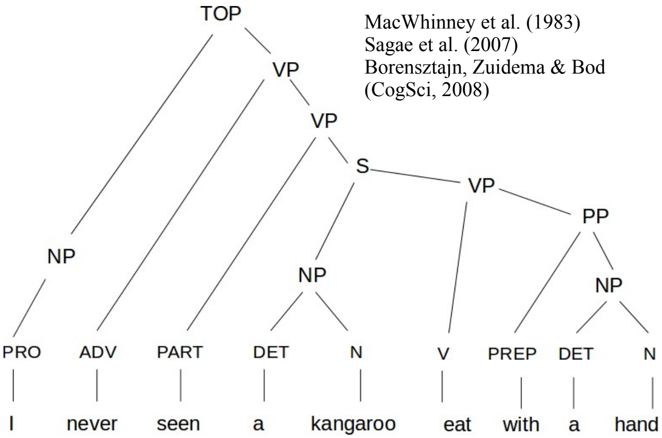
Jelle Zuidema
ILLC, Universiteit van Amsterdam

Unsupervised Language Learning, 2014

2;5 *CHI: seen one those .
3;0 *CHI: I never seen a watch .
3;0 *CHI: I never seen a watch .
3;0 *CHI: I never seen a bandana .
3;0 *CHI: I never seen a monkey train .
3;0 *CHI: I never seen a tree dance .
3;2 *CHI: I never seen a duck like that # riding@o on a pony .
3;2 *CHI: I never seen (a)bout dat [: that] .
3;5 *CHI: I never seen this jet .
3;5 *CHI: I never seen this jet .
3;5 *CHI: I never seen a Sky_Dart .
3;5 *CHI: I never seen this before .
3;8 *CHI: yeah # I seen carpenters too .
3;8 *CHI: where had you seen carpenters do that ?
3;8 *CHI: I never seen her .
3;8 *CHI: I never seen people wear de [: the] fish flies .
3;8 *CHI: where have you seen a whale ?
3;8 *CHI: I never seen a bird talk .
3;11 *CHI: I never seen a kangaroo knit .
3;11 *CHI: I never seen dat [: that] to play .
3;11 *CHI: I never seen a dog play a piano # have you ?
3;11 *CHI: I never seen a rhinoceros eat with a hands .
4;7 *CHI: I seen one in the store some days .

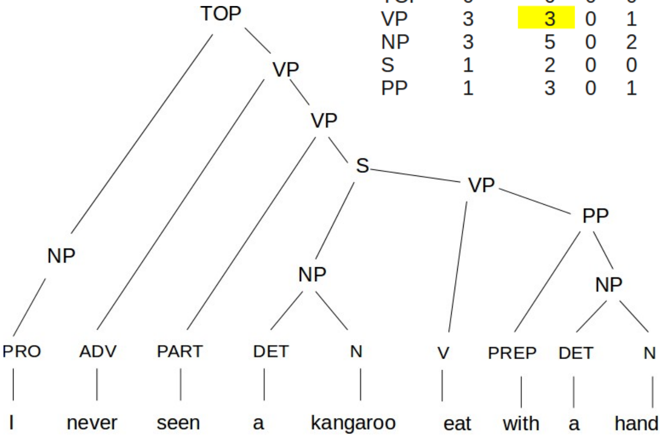
Brown corpora in Childes

MacWhinney et al. (1983)
Sagae et al. (2007)
Borensztajn, Zuidema & Bod (CogSci, 2008)



Adam, 3;11.01

	TOP	VP	NP	S	PP
TOP	0	0	0	0	0
VP	3	3	0	1	0
NP	3	5	0	2	1
S	1	2	0	0	0
PP	1	3	0	1	0



Adam, 3;11.01

Human vs. Animal Communication

- Border collie Rico (Kaminski et al. 2004, Science): 200 names for objects.
- Human infants (O'Grady, 2005): estimated to learn 10 words a day between age 2 and 6 (≈14,000 words at age 6, ≈60,000 words at age 12)
- Bonobo Kanzi (Savage-Rumbaugh & Lewin, 1994; Truswell, *in prep.*): knows the meanings of dozens of content words (verbs and nouns), but scores at chance when tested with coordination structure as in "Put the coke and the milk in the fridge".

Time flies like an arrow.
الوقت الذباب يشبه السهم.
time-DEF flies-DEF resemble arrow-DEF
"Time flies like arrow."

Fruit flies like a banana.
ذباب الفاكهة مثل الموز.
Flies-N fruit-N resemble banana-N
"Fruit flies like bananas."

Steedman, 2008, *CL*

This is the bank that bought the company.

وهذا البنك هو ان اشترت الشركة.

"This is the bank that bought the company."

This is the company that the bank bought.

هذه هي الشركة التي اشترت البنك.

"*This is the company that bought the bank."

This is the bank that wants to buy the company.

هذا هو المصرف الذي يريد لشراء الشركة.

"This is the bank, which wants to buy the company."

This is the company which the bank wants to buy.

هذه هي الشركة التي تريد شراء البنك.

"*This is the company that wants to buy the bank."

This is the company that said the bank bought bonds.

هذه هي الشركة التي قال البنك بشراء السندات.

"This is the company that said the bank bought the bonds."

This is the company that the bank said bought bonds.

هذه هي الشركة التي قال البنك بشراء السندات.

"*This is the company that said the bank bought the bonds."

These are the bonds that the company said that the bank bought.

هذه هي سندات الشركة أن البنك اشترى.

"*These are the bonds that the bank bought the company."

Grammar in other NLP domains

- E.g., speech recognition
  - please, right this down
  - write now
  - who's write, and who's wrong
- E.g., anaphora resolution
  - Mary didn't know who John was married to. He told her, and it turned out, she already knew her.
- E.g., machine translation

Annotated/unannotated data

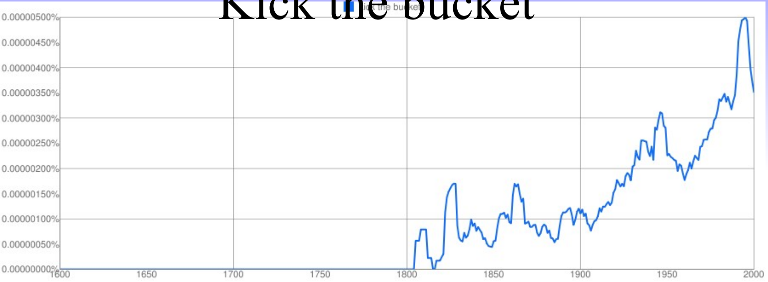
- Syntactically Annotated corpora
  - Penn WSJ Treebank
    trainset: 38k sentences, ~1M words
  - Tuebingen spoken/written English/German
  - Corpus Gesproken Nederlands
- Unannotated corpora
  - the web ...
  - Google's ngram corpora

Spam

www.culturomics.org

Penn WSJ: 0 counts.

Kick the bucket



www.culturomics.org

Penn WSJ: 0 counts.

... know but were afraid to ...

www.culturomics.org

Penn WSJ: 0 counts.

This course (I)

- We discuss the recent advances in computational methods to learn language from unlabeled data;
- Focus on grammar;
- Relevance for cognitive science and language technology in the back of our minds.

This course (II)

- Part of the Natural Language Processing and Learning track of the MSc A.I.; optional course for students in Logic, Brain & Cognitive Science;
- Introduction to current research; setup of a research seminar
- Requires active participation!

This course (III)

- 8 reading assignments (8 supportive lectures - presence required)
  - For Thursday, read Griffiths & Yuille (2007)
- 4 practical assignments (8 supportive lab sessions - presence recommended)
- Miniproject (week 5-7)
- Workshop with presentations (exam week, 24 & 25/3, 13-17h - or some other day - presence required)

This course (IV)

Assessment

- Forum posts & discussions (25%) - come prepared to class!
- Practical assignments (25%)
- Final report & presentation (50%; literature review 20%, practical work 30%)

Formalisms to describe language

Chomsky finds Finite-state Automata inadequate

(Chomsky, 1957)

Let S_1, S_2, S_3, S_4 be simple declarative sentences in English.

1. If S_1, then S_2.
2. Either S_3 or S_4.
3. The man who said that S_5, is arriving today

Extended Chomsky Hierarchy

language	grammar	automaton
	Set	Look-up table
Locally testable	ngram	
Regular	Left-linear	Finite-state
Context-free	Context-free	Pushdown
Tree-Adjoining	Tree-Adjoining	Embedded pushdown
Mildly context-sens.	Range Concatenation	Thread
Recursively enum.	Unrestricted	Turing Machine

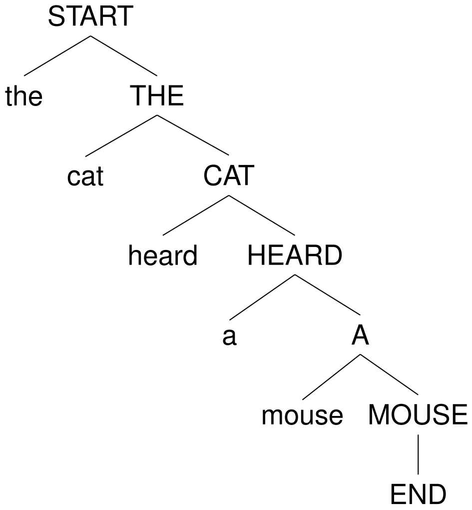
Extended Chomsky Hierarchy

language	grammar	rules
$\{a, b, cbabb\}$	Set	$\in$
$(ab)^n$	ngram	$\langle a, b\rangle, \langle b, a\rangle, \langle ab, a\rangle$
$a^n ba^m$	Left-linear	$S \rightarrow AB, B \rightarrow bA$
$a^n b^n$	Context-free	$S \rightarrow aSb, S \rightarrow ab$
$a^n b^n c^n d^n \vert 1 \leq n$	Tree-Adjoining	
	Range Concatenation	$S[abc] \rightarrow A[a, c]B[b]$
	Unrestricted	

Probabilistic Extensions

grammar	probabilistic grammar
Set	Probability distribution
ngram	Markov model
Left-linear	Hidden Markov (HMM)
Context-free	PCFG
Tree-Adjoining	PTAG
Range Concatenation	PLCRS
Unrestricted	

Relation between PCFGs, HMMs and Ngrams



Relation between PCFGs, HMMs and Ngrams

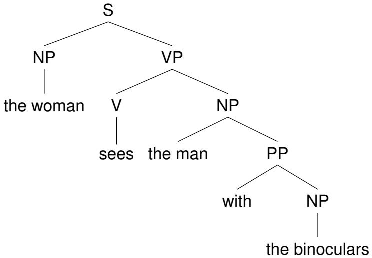
- Class of ngrams is proper subset of class of HMMs
- Class of HMMs is proper subset of class of PCFGs.
- Ngrams can be implemented as PCFGs with nonterminals corresponding to history of length $n - 1$
- rules are right-linear, trees produced are right-branching.

Probabilistic Context Free Grammars

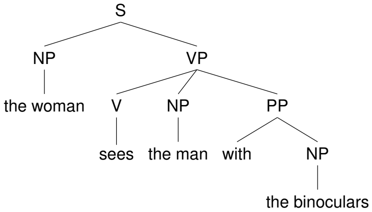
s: The woman seesthe man with the binoculars

G:

S	$\rightarrow$ NP VP	1.0
NP	$\rightarrow$ the woman	0.2
NP	$\rightarrow$ the man	0.2
NP	$\rightarrow$ the binoculars	0.2
NP	$\rightarrow$ the dress	0.2
NP	$\rightarrow$ NP PP	0.2
VP	$\rightarrow$ V NP	0.7
VP	$\rightarrow$ V NP PP	0.3
PP	$\rightarrow$ with NP	1.0
V	$\rightarrow$ sees	1.0



$d_1(s)$, $p = .7 \times .2 \times .2 = .028$



$d_2(s)$, $p = .3 \times .2 \times .2 = .012$

Conditioning context & History

Conditioning context & History

History-based Grammars

S→NP VP
NP→it
NP→the dog
VP→V NP
V→chases

event	conditioning context
NP VP	S
the dog	NP
V NP	VP
chases	V
it	NP

Conditioning context & History

History-based Grammars

S→NP VP
NP→it
NP→the dog
VP→V NP
V→chases

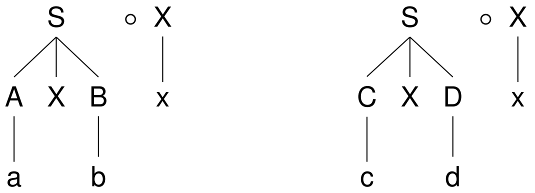
event	conditioning context
NP VP	S, BEGIN
the dog	NP, S
V NP	VP, S
chases	V, VP
it	NP, VP

Commitment

Commitment

Consider the language YXY’, where $Y, Y' \in \{a, b, c, d\}$ and Y predicts Y'.

Tree Substitution Grammars



Computational Tasks

Recognition $s \in L_G = \{s | \exists d (y(d) = s) \;\; \& \;\; \forall r \in d (r \in G)\}$

Parsing $D(s) = \{d | y(d) = s \;\; \& \;\; \forall r \in d (r \in G)\}$

Disambiguation $\arg\max_{d \in D(s)} P(d) = \arg\max_{d \in D(s)} \prod_{r \in d} P(r)$

Inference $\arg\max_{\vec{p}} P(\vec{s} | G, \vec{p})$

Computational Tasks

Recognition Earley's and CYK algorithms: charts with function returning true/false

Parsing function returning list of partial parse

Disambiguation Viterbi: functioning returning maximum probability (and partial parse)

Inference Expectation-Maximization (Thursday)

Transducers / Synchronous Grammars

grammar	transducer	
Set	Map	$\langle m_1, w_1 \rangle, \langle m_2, w_2 \rangle, \ldots$
ngram		
Left-linear	Finite-state transducer	
Context-free	SCFG	
Tree-Adjoining	STAG	
Range Concatenation		
Unrestricted	(lambda calculus)	

A very brief tour of statistical learning

Bayes' Rule

$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$

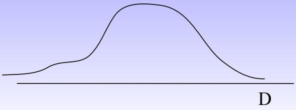
Bayes' Rule

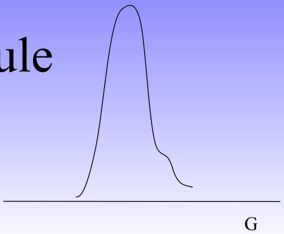
likelihood *prior*

$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$

posterior *probability of data*

Bayes' Rule

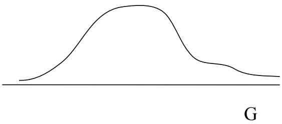




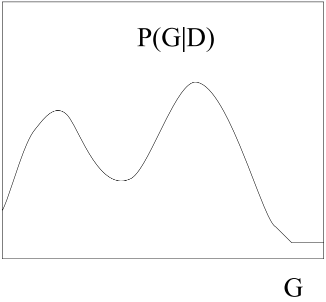
likelihood *prior*

$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$

posterior *probability of data*

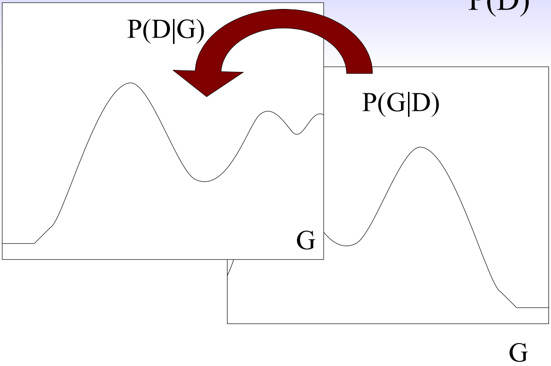


Statistical inference



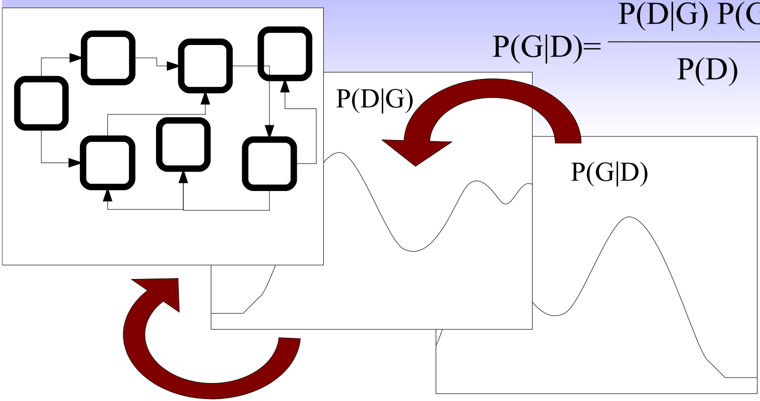
Statistical inference

$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$

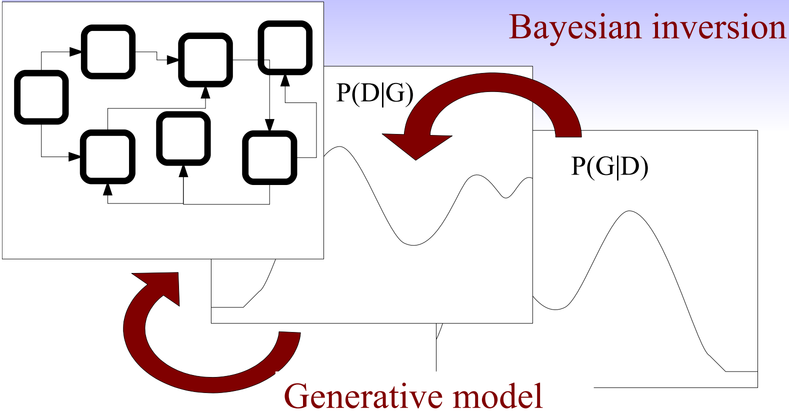


Statistical inference

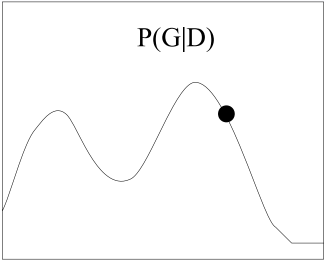
$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$



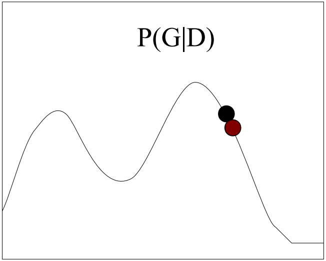
Statistical inference



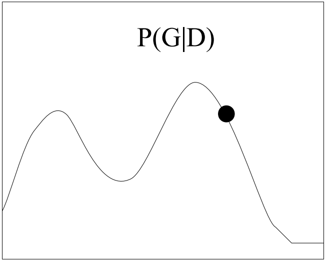
Stochastic hillclimbing



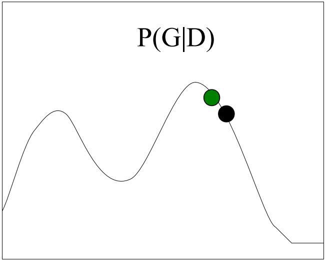
Stochastic hillclimbing



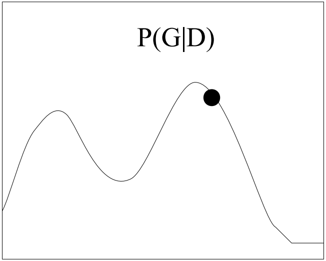
Stochastic hillclimbing



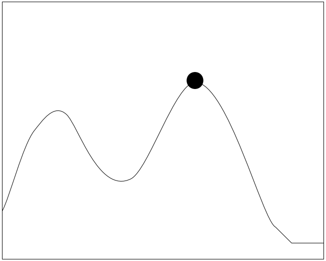
Stochastic hillclimbing



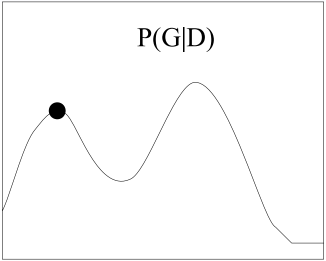
Stochastic hillclimbing



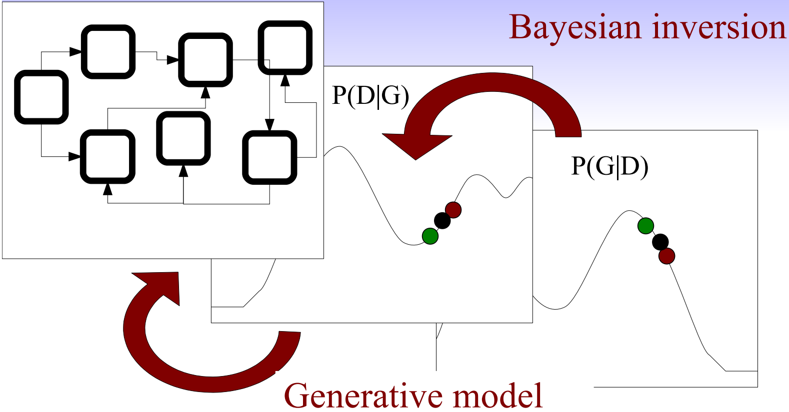
Stochastic hillclimbing



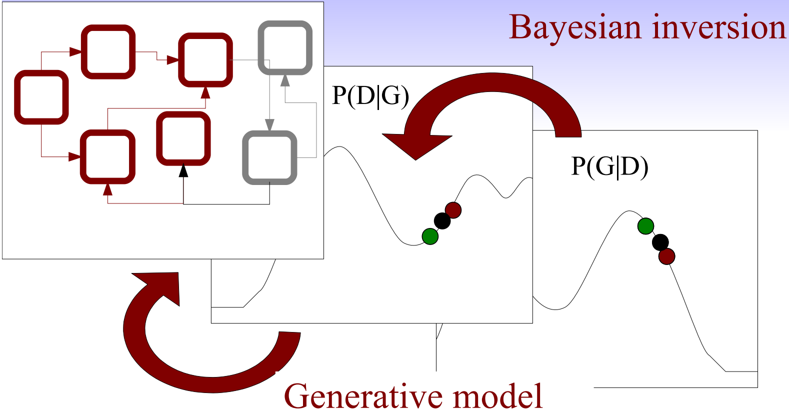
Local optimum



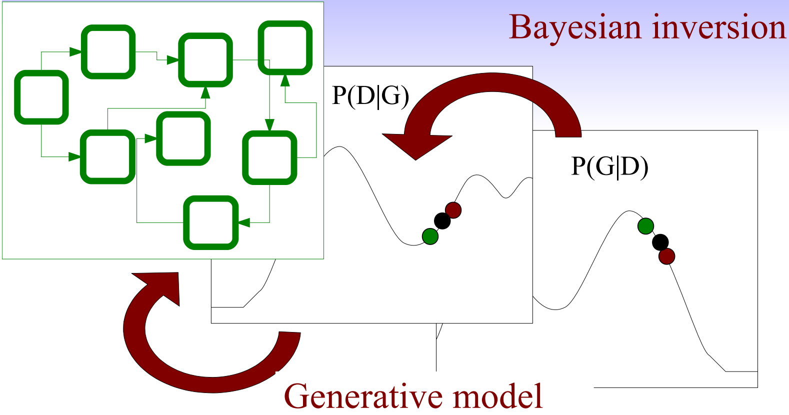
Statistical inference



Statistical inference



Statistical inference



- Generative models can be used to define probability distributions over possibly infinite sets of complex structures (e.g. phrase-structure trees);
- Bayesian inversion allows us to express the posterior $P(G|D)$ in terms of the likelihood $P(D|G)$ and prior $P(G)$.
- If the prior is uniform, the most probable posterior (MAP) grammar is also the maximum likelihood grammar (ML).