Constituency Parsing

Computational Linguistics: Jordan Boyd-Graber
University of Maryland
INTRO / CHART PARSING

A More Grounded Syntax Theory

- A central question in linguistics is **how do we know when a sentence is grammatical**?
- Chomsky's generative grammars attempted to mathematically formalize this question
- Linguistic phrases contained a universal, hierarchical structure formalized as parse trees

A More Grounded Syntax Theory

- A central question in linguistics is **how do we know when a sentence is grammatical**?
- Chomsky's generative grammars attempted to mathematically formalize this question
- Linguistic phrases contained a universal, hierarchical structure formalized as parse trees
- Today
  - A formalization
  - Foundation of all computational syntax
  - Learnable from data

Context Free Grammars

Definition

- N: finite set of non-terminal symbols
- Σ: finite set of terminal symbols
- R: productions of the form $X \to Y_1 \dots Y_n$, where $X \in N$, $Y \in (N \cup \Sigma)$
- S: a start symbol within N

Examples of non-terminals:

- np for "noun phrase"
- vp for "verb phrase"
- Often correspond to multiword syntactic abstractions

Context Free Grammars

Definition

- N: finite set of non-terminal symbols
- Σ: finite set of terminal symbols
- R: productions of the form $X \rightarrow Y_1 \ldots Y_n$, where $X \in N$, $Y \in (N \cup \Sigma)$
- S: a start symbol within N

Examples of terminals:

- *"dog"*
- *"play"*
- *"the"*

Context Free Grammars

Definition

- N: finite set of non-terminal symbols
- Σ: finite set of terminal symbols
- R: productions of the form $X \longrightarrow Y_1 \ldots Y_n$, where $X \in N$, $Y \in (N \cup \Sigma)$
- S: a start symbol within N

Examples of productions:

- n $\longrightarrow$ "*dog*"
- np $\longrightarrow$ n
- np $\longrightarrow$ adj n

Context Free Grammars

Definition

- N: finite set of non-terminal symbols
- Σ: finite set of terminal symbols
- R: productions of the form $X \longrightarrow Y_1 \dots Y_n$, where $X \in N$, $Y \in (N \cup \Sigma)$
- S: a start symbol within N

In NLP applications, by convention we use S as the start symbol

Flexibility of CFG Productions

- Unary rules: nn → "*man*"
- Mixing terminals and nonterminals on RHS:
  - np → "*Congress*" Vt "*the*" "*pooch*"
  - np → "*the*" nn
- Empty terminals
  - np → ϵ
  - adj → ϵ

Derivations

- A derivation is a sequence of strings $s_1 \dots s_T$ where
- $s_1 \equiv S$, the start symbol
- $s_T \in \Sigma^*$: i.e., the final string is only terminals
- $s_i, \forall i > 1$, is derived from s_{i-1} by replacing some non-terminal X in s_{i-1} and replacing it by some β, where $x \rightarrow \beta \in R$.

Derivations

- A derivation is a sequence of strings $s_1 \dots s_T$ where
- $s_1 \equiv S$, the start symbol
- $s_T \in \Sigma^*$: i.e., the final string is only terminals
- $s_i, \forall i > 1$, is derived from s_{i-1} by replacing some non-terminal X in s_{i-1} and replacing it by some β, where $x \rightarrow \beta \in R$.
- Example: parse tree

Example Derivation

Productions

s → np vp	np → Det nn	vp → vz
vp → AdvP vz	np → AdjP nn	np → pro
Det → *"the"*	Det → *"a"*	Det → *"an"*
nn → *"dot"*	nn → *"cat"*	nn → *"mouse"*
vz → *"barked"*	vz → *"ran"*	vz → *"sat"*
⋮	⋮	⋮

$s_1 =$

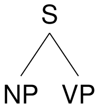
S

Example Derivation

Productions

s → np vp	np → Det nn	vp → vz
vp → AdvP vz	np → AdjP nn	np → pro
Det → "*the*"	Det → "*a*"	Det → "*an*"
nn → "*dot*"	nn → "*cat*"	nn → "*mouse*"
vz → "*barked*"	vz → "*ran*"	vz → "*sat*"
⋮	⋮	⋮

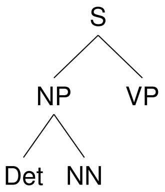
$s_2 =$



Productions

s → np vp	np → Det nn	vp → vz
vp → AdvP vz	np → AdjP nn	np → pro
Det → "*the*"	Det → "*a*"	Det → "*an*"
nn → "*dot*"	nn → "*cat*"	nn → "*mouse*"
vz → "*barked*"	vz → "*ran*"	vz → "*sat*"
⋮	⋮	⋮

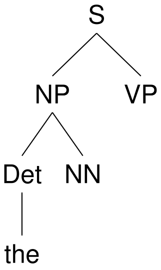
$s_3 =$



Productions

s → np vp	np → Det nn	vp → vz
vp → AdvP vz	np → AdjP nn	np → pro
Det → “*the*”	Det → “*a*”	Det → “*an*”
nn → “*dot*”	nn → “*cat*”	nn → “*mouse*”
vz → “*barked*”	vz → “*ran*”	vz → “*sat*”
⋮	⋮	⋮

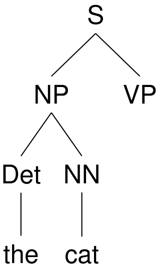
$s_4 =$



Productions

s → np vp	np → Det nn	vp → vz
vp → AdvP vz	np → AdjP nn	np → pro
Det → "*the*"	Det → "*a*"	Det → "*an*"
nn → "*dot*"	nn → "*cat*"	nn → "*mouse*"
vz → "*barked*"	vz → "*ran*"	vz → "*sat*"
⋮	⋮	⋮

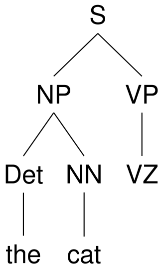
$s_5 =$



Productions

s → np vp	np → Det nn	vp → vz
vp → AdvP vz	np → AdjP nn	np → pro
Det → *"the"*	Det → *"a"*	Det → *"an"*
nn → *"dot"*	nn → *"cat"*	nn → *"mouse"*
vz → *"barked"*	vz → *"ran"*	vz → *"sat"*
⋮	⋮	⋮

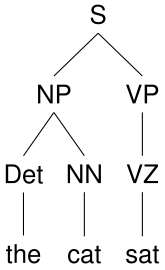
$s_6 =$



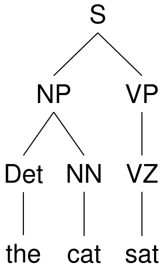
Productions

s → np vp	np → Det nn	vp → vz
vp → AdvP vz	np → AdjP nn	np → pro
Det → *"the"*	Det → *"a"*	Det → *"an"*
nn → *"dot"*	nn → *"cat"*	nn → *"mouse"*
vz → *"barked"*	vz → *"ran"*	vz → *"sat"*
⋮	⋮	⋮

$s_7 =$



Example Derivation



Ambiguous Yields

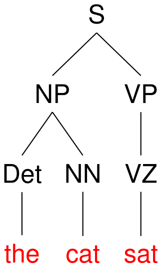
The **yield** of a parse tree is the collection of terminals produced by the parse tree. Given a yield s.

Parsing / Decoding

Given, a yield s and a grammar G, determine the set of parse trees that could have produced that sequence of terminals: $T_G(s)$.

$S_7 =$

Example Derivation



Ambiguous Yields

The **yield** of a parse tree is the collection of terminals produced by the parse tree. Given a yield s.

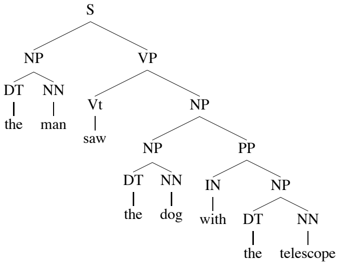
Parsing / Decoding

Given, a yield s and a grammar G, determine the set of parse trees that could have produced that sequence of terminals: $T_G(s)$.

Ambiguity

Example sentence: "The man saw the dog with the telescope"

- Grammatical: $T_G(s) > 0$
- Ambiguous: $T_G(s) > 1$

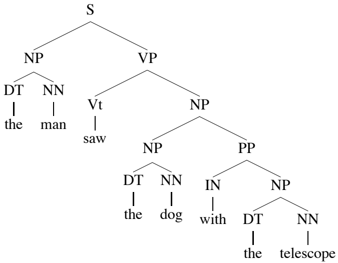


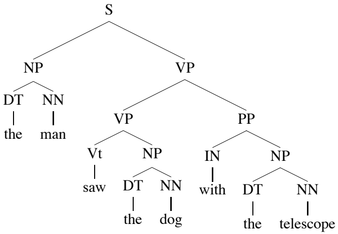
- Which should we prefer?

Ambiguity

Example sentence: "The man saw the dog with the telescope"

- Grammatical: $T_G(s) > 0$
- Ambiguous: $T_G(s) > 1$





- Which should we prefer?
- One is more probable than the other
- Add probabilities!

Goals

- What we want is a probability distribution over possible parse trees $t \in T_G(s)$

$$\forall t, p(t) \geq 0 \qquad \sum_{t \in T_G(s)} p(t) = 1 \tag{1}$$

- Rest of this lecture:
  - How do we define the function $p(t)$ (paramterization)
  - How do we learn $p(t)$ from data (estimation)
  - Given a sentence, how do we find the possible parse trees (parsing / decoding)

Parametrization

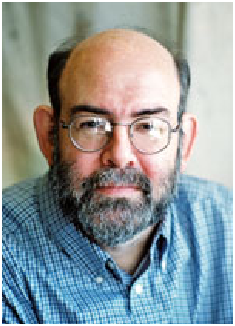
- For every production $\alpha \rightarrow \beta$, we assume we have a function $q(\alpha \rightarrow \beta)$
- We consider it a **conditional probability** of β (LHS) being derived from α (RHS)

$$\sum_{\alpha \rightarrow \beta \in R : \alpha = X} q(\alpha \rightarrow \beta) = 1 \quad (2)$$

- The total probability of a tree $t \equiv \{\alpha_1 \rightarrow \beta_1 \ldots \alpha_n \rightarrow \beta_n\}$ is

$$p(t) = \prod_{i=1}^{n} q(\alpha_i \rightarrow \beta_i) \quad (3)$$

Estimation

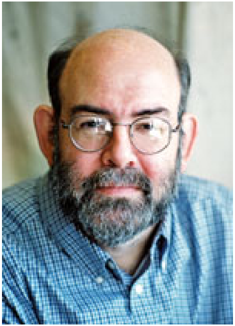


- Get a bunch of grad students to make parse trees for a million sentences
- Mitch Markus: Penn Treebank (Wall Street Journal)
- To compute the conditional probability of a rule,

$$q(\text{np} \rightarrow \text{Det} \quad \text{adj} \quad \text{nn}) \approx \frac{\text{Count}(\text{np} \rightarrow \text{Det} \quad \text{adj} \quad \text{nn})}{\text{Count}(\text{np})}$$

- Where “Count” is the number of times that derivation appears in the sentences

Estimation



- Get a bunch of grad students to make parse trees for a million sentences
- Mitch Markus: Penn Treebank (Wall Street Journal)
- To compute the conditional probability of a rule,

$$q(\text{np} \rightarrow \text{Det} \quad \text{adj} \quad \text{nn}) \approx \frac{\text{Count}(\text{np} \rightarrow \text{Det} \quad \text{adj} \quad \text{nn})}{\text{Count}(\text{np})}$$

- Where "Count" is the number of times that derivation appears in the sentences
- Why no smoothing?

Dynamic Programming

- Like for dependency parsing, we build a chart to consider all possible subtrees
- First, however, we'll just consider whether a sentence is grammatical or not
- Build up a chart with all possible derivations of spans
- Then see entry with start symbol over the entire sentence: those are all grammatical parses

CYK Algorithm (deterministic)

Assumptions

Assumes binary grammar (not too difficult to extend) and no recursive rules

Given sentence $\vec{w}$ of length N, grammar (N, Σ, R, S)
Initialize array $C[s, t, n]$ as array of booleans, all false ($\perp$)

CYK Algorithm (deterministic)

Assumptions

Assumes binary grammar (not too difficult to extend) and no recursive rules

Given sentence $\vec{w}$ of length N, grammar (N, Σ, R, S)
Initialize array $C[s,t,n]$ as array of booleans, all false ($\perp$)
for $i = 0 \ldots N$ **do**
  for For each production $r_j \equiv N_a \longrightarrow w_i$ **do**
    set $C[i,i,a] \leftarrow \top$

CYK Algorithm (deterministic)

Assumptions

Assumes binary grammar (not too difficult to extend) and no recursive rules

Given sentence $\vec{w}$ of length N, grammar (N, Σ, R, S)
Initialize array $C[s, t, n]$ as array of booleans, all false ($\perp$)
for $i = 0 \ldots N$ **do**
    for For each production $r_j \equiv N_a \longrightarrow w_i$ **do**
        set $C[i, i, a] \leftarrow \top$
for $l = 2 \ldots n$ (length of span) **do**
    for $s = 1 \ldots N - l + 1$ (start of span) **do**
        for $k = 1 \ldots l - 1$ (pivot within span) **do**
            for each production $r \equiv \alpha \longrightarrow \beta\gamma$ **do**
                if $\neg C[s, s + l, \alpha]$ **then**
                    $C[s, s + l, \alpha] \leftarrow C[s, s + k - 1, \beta] \wedge C[s + k, s + l, \gamma]$

Chart Parsing

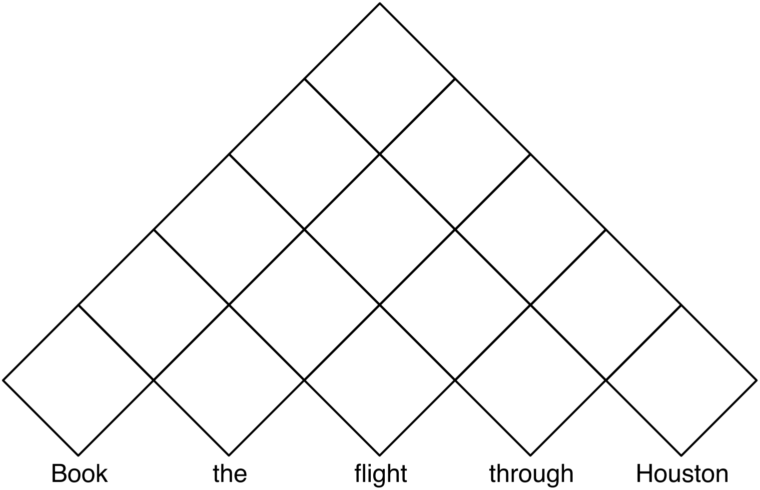


Chart Parsing

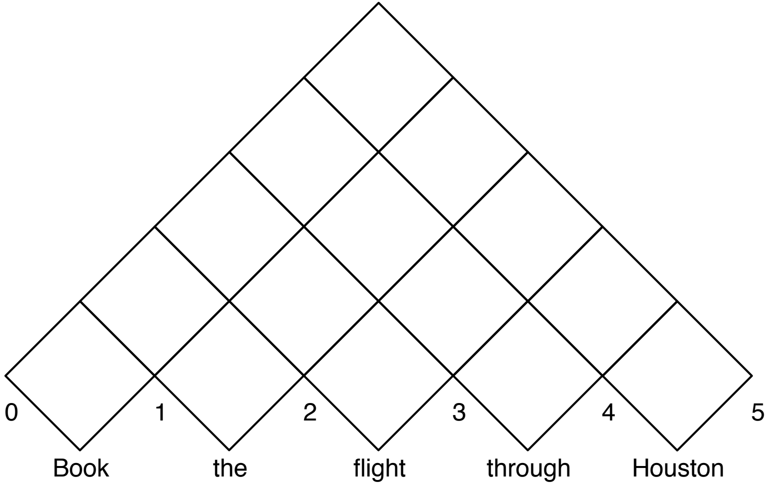


Chart Parsing

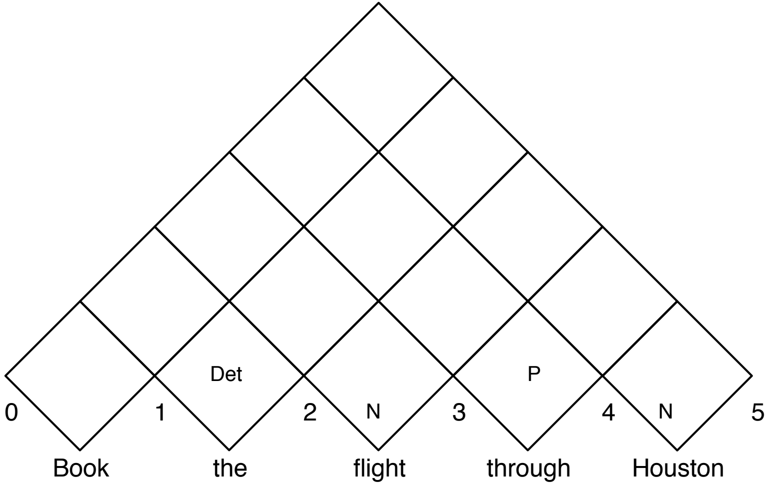


Chart Parsing

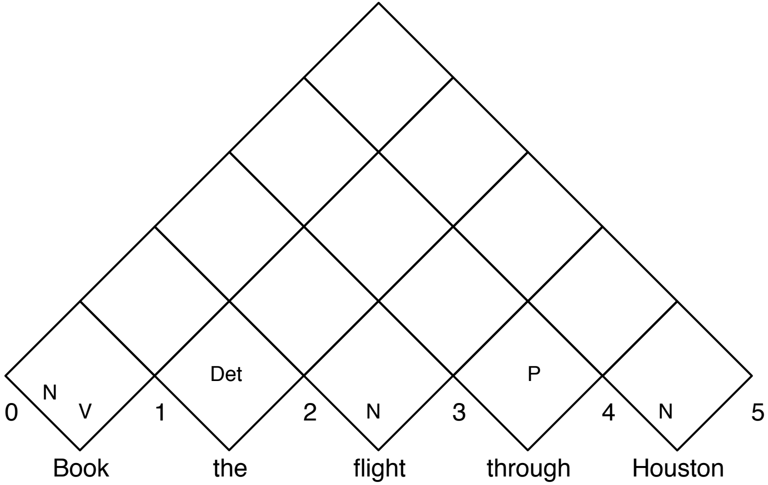


Chart Parsing

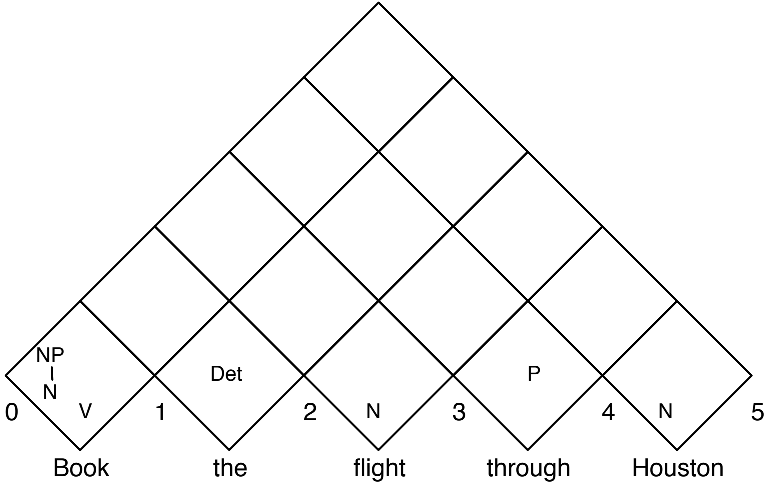


Chart Parsing

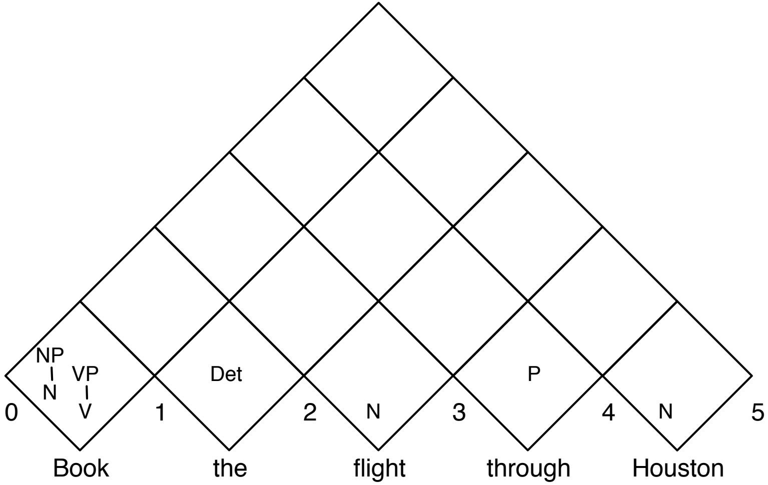


Chart Parsing

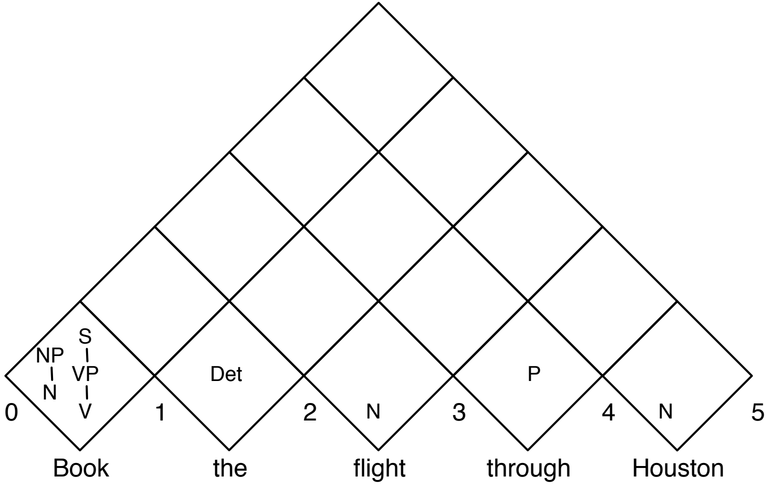


Chart Parsing

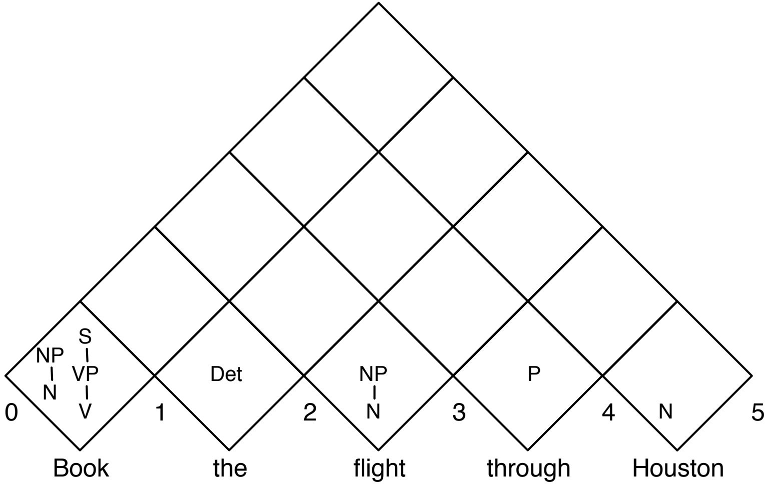


Chart Parsing

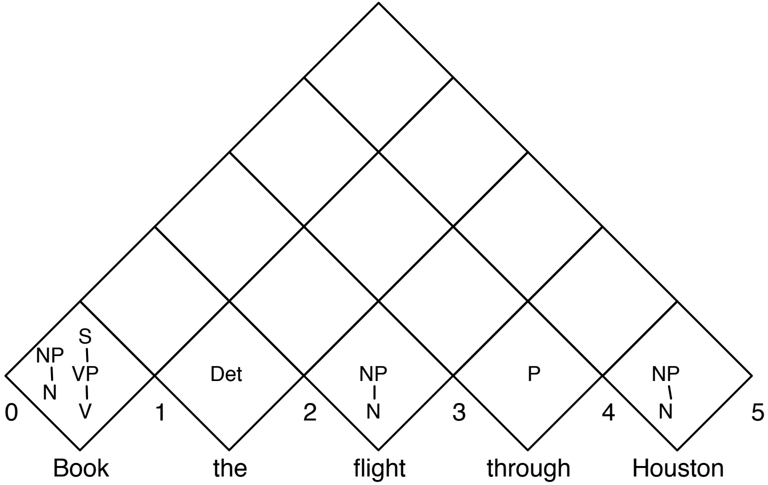


Chart Parsing

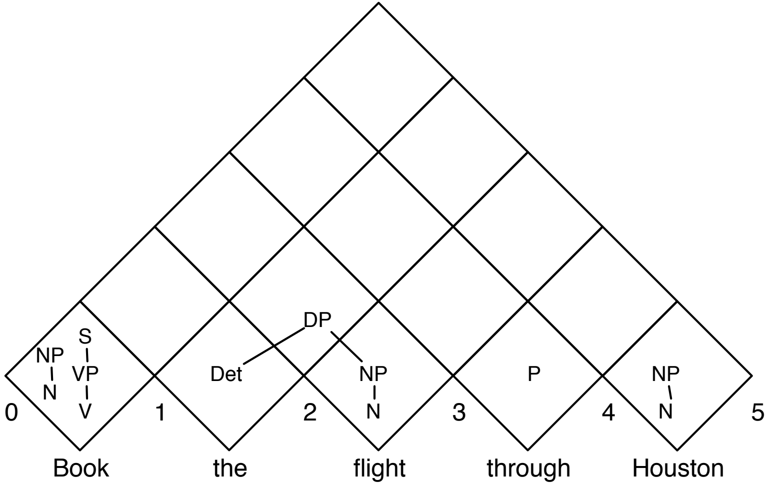


Chart Parsing

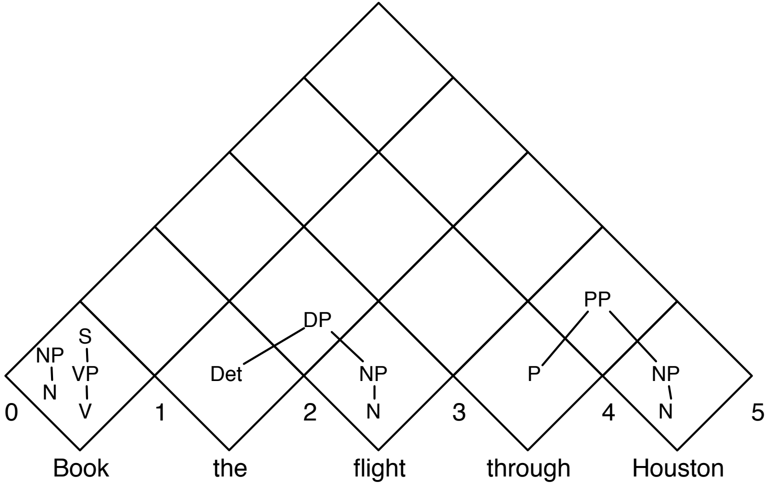


Chart Parsing

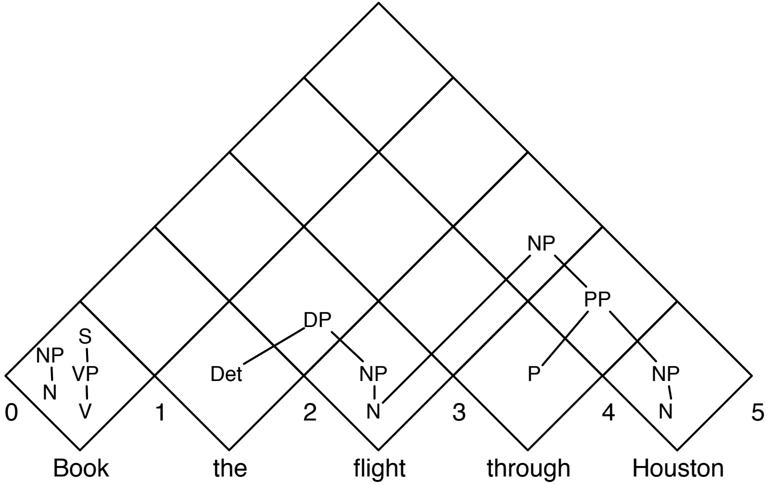


Chart Parsing

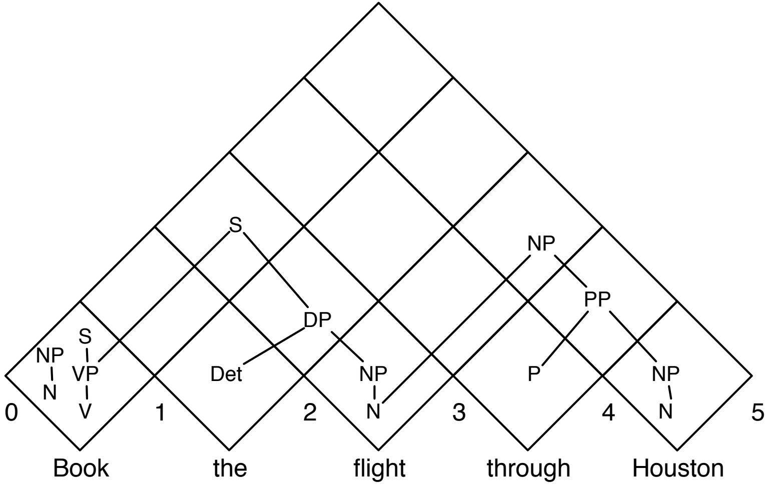


Chart Parsing

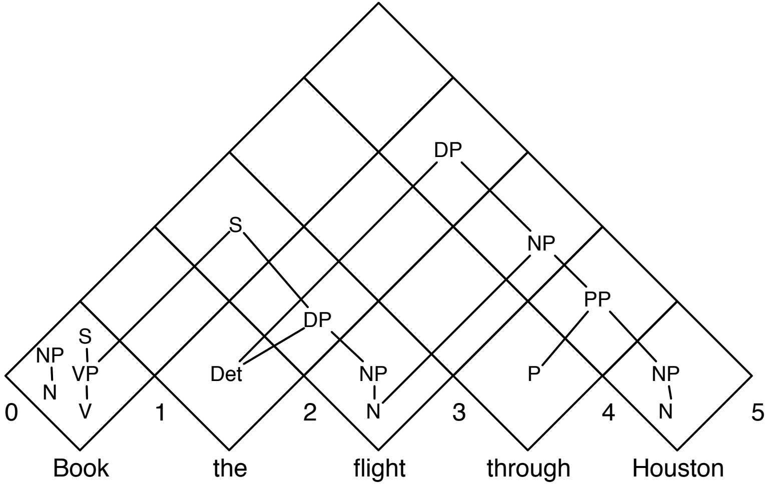
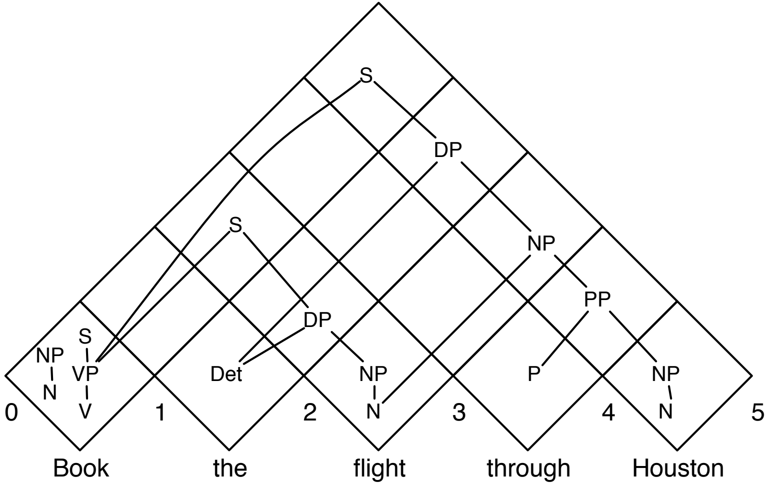


Chart Parsing



Complexity?

- Chart has n^2 cells

Complexity?

- Chart has n^2 cells
- Each cell has n options

Complexity?

- Chart has n^2 cells
- Each cell has n options
- Times the number of productions $|G|$

Complexity?

- Chart has n^2 cells
- Each cell has n options
- Times the number of productions $|G|$
- Thus, $O(n^3|G|)$

How to deal with PCFG ambiguity

- In addition to keeping track of non-terminals in cell, also include **max** probability of forming non-terminal from sub-trees

  $$C[s,s+k,\alpha] \leftarrow \max(C[s,s+k,\alpha], C[s,s+l-1,\beta] \cdot C[s+l,s+k,\gamma])$$

- The score associated with S in the top of the chart is the best overall parse-tree (**given the yield**)

Recap

- Hierarchical syntax model: context free grammar
- Probabilistic interpretation: learn from data to solve ambiguity
- In class (next time):
  - Work through example to resolve ambiguity
  - Scoring a sentence

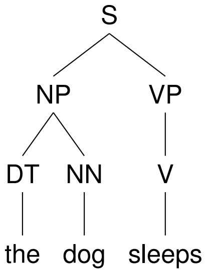
A pcfg

Assume the following grammar

s	⟶	np	vp	1.0	v	⟶	sleeps	0.4
vp	⟶	v	np	0.7	v	⟶	saw	0.6
vp	⟶	vp	pp	0.2	nn	⟶	man	0.1
vp	⟶	v		0.1	nn	⟶	woman	0.1
np	⟶	dt	nn	0.2	nn	⟶	telescope	0.3
np	⟶	np	pp	0.8	nn	⟶	dog	0.5
pp	⟶	p	np	1.0	dt	⟶	the	1.0
					p	⟶	with	0.6
					p	⟶	in	0.4

Evaluating the probability of a sentence

What is the probability of the parse



Evaluating the probability of a sentence

$$\underbrace{1.0}_{\text{det} \longrightarrow \text{the}} \cdot \underbrace{0.5}_{\text{n} \longrightarrow \text{dog}} \cdot \underbrace{1.0}_{\text{v} \longrightarrow \text{sleeps}} \cdot \underbrace{0.1}_{\text{vp} \longrightarrow \text{v}} \cdot \underbrace{0.2}_{\text{np} \longrightarrow \text{dt n}} \cdot \underbrace{1.0}_{\text{s} \longrightarrow \text{np vp}} = 0.002$$

Span 0

1. $C[8,8,\text{nn}] = \ln(0.3) = -1.2$
2. $C[7,7,\text{dt}] = \ln(1.0) = 0.0$
3. $C[6,6,\text{p}] = \ln(0.6) = -0.51$
4. $C[5,5,\text{nn}] = \ln(0.5) = -0.69$
5. $C[4,4,\text{dt}] = \ln(1.0) = 0.0$
6. $C[3,3,\text{v}] = \ln(0.6) = -.51$
7. $C[3,3,\text{vp}] = \ln(0.6) + \ln(0.1) = -2.8$
8. $C[2,2,\text{nn}] = \ln(0.1) = -2.3$
9. $C[1,1,\text{dt}] = \ln(1.0) = 0.0$

Span 1

1. $C[1,2,\text{np}] = \underbrace{0.0}_{\text{C[1,1,DT]}} + \ln(\underbrace{-2.3}_{\text{C[2,2,NN]}}) + \ln(\underbrace{0.2}_{\text{np} \rightarrow \text{dt n}}) = -2.3 + -1.6 = -3.9$

Span 1

1. $C[1,2,\text{np}] = \underbrace{0.0}_{\text{C[1,1,DT]}} + \ln(\underbrace{-2.3}_{\text{C[2,2,NN]}}) + \ln(\underbrace{0.2}_{\text{np} \rightarrow \text{dt n}}) = -2.3 + -1.6 = -3.9$

2. $C[4,5,\text{np}] = \underbrace{0.0}_{\text{C[4,4,DT]}} + \underbrace{-.69}_{\text{C[5,5,NN]}} + \ln(\underbrace{0.2}_{\text{np} \rightarrow \text{dt n}}) = -0.69 + -1.6 = -2.3$

Span 1

1. $C[1,2,\text{np}] = \underbrace{0.0}_{\text{C[1,1,DT]}} + \ln(\underbrace{-2.3}_{\text{C[2,2,NN]}}) + \ln(\underbrace{0.2}_{\text{np} \rightarrow \text{dt n}}) = -2.3 + -1.6 = -3.9$

2. $C[4,5,\text{np}] = \underbrace{0.0}_{\text{C[4,4,DT]}} + \underbrace{-.69}_{\text{C[5,5,NN]}} + \ln(\underbrace{0.2}_{\text{np} \rightarrow \text{dt n}}) = -0.69 + -1.6 = -2.3$

3. $C[7,8,\text{np}] = \underbrace{0.0}_{\text{C[7,7,DT]}} + \underbrace{-1.2}_{\text{C[8,8,NN]}} + \ln(\underbrace{0.2}_{\text{np} \rightarrow \text{dt n}}) = -1.2 + -1.6 = -2.8$

Span 2

1. $C[1,3,s] = \underbrace{-3.9}_{\text{C[1,2,NP]}} + \underbrace{-2.8}_{\text{C[3,3,VP]}} + \ln(\underbrace{1.0}_{\text{s} \rightarrow \text{np vp}}) = -6.7$

Span 2

1. $C[1,3,\mathrm{s}] = \underbrace{-3.9}_{\text{C[1,2,NP]}} + \underbrace{-2.8}_{\text{C[3,3,VP]}} + \ln(\underbrace{1.0}_{\text{s} \rightarrow \text{np vp}}) = -6.7$
2. $C[3,5,\mathrm{vp}] = \underbrace{-0.5}_{\text{C[3,3,V]}} + \underbrace{-2.3}_{\text{C[4,5,NP]}} + \ln(\underbrace{0.7}_{\text{vp} \rightarrow \text{v np}}) = -2.8 - 0.36 = -3.2$

Span 2

1. $C[1,3,\text{s}] = \underbrace{-3.9}_{\text{C[1,2,NP]}} + \underbrace{-2.8}_{\text{C[3,3,VP]}} + \ln(\underbrace{1.0}_{\text{s} \rightarrow \text{np vp}}) = -6.7$
2. $C[3,5,\text{vp}] = \underbrace{-0.5}_{\text{C[3,3,V]}} + \underbrace{-2.3}_{\text{C[4,5,NP]}} + \ln(\underbrace{0.7}_{\text{vp} \rightarrow \text{v np}}) = -2.8 - 0.36 = -3.2$
3. $C[6,8,\text{pp}] = \underbrace{-0.51}_{\text{C[6,6,P]}} + \underbrace{-2.8}_{\text{C[7,8,NP]}} + \ln(\underbrace{1.0}_{\text{pp} \rightarrow \text{p np}}) = -3.3 + -1.6 = -3.3$

Span 4

1. $C[1,5,s] = \underbrace{-3.9}_{\text{C[1,2,NP]}} + \underbrace{-3.2}_{\text{C[3,5,VP]}} + \ln(\underbrace{1.0}_{\text{s} \rightarrow \text{np vp}}) = -7.1$

Span 4

1. $C[1,5,\text{s}] = \underbrace{-3.9}_{\text{C[1,2,NP]}} + \underbrace{-3.2}_{\text{C[3,5,VP]}} + \ln(\underbrace{1.0}_{\text{s} \rightarrow \text{np vp}}) = -7.1$
2. $C[4,8,\text{np}] = \underbrace{-2.3}_{\text{C[4,5,NP]}} + \underbrace{-3.3}_{\text{C[6,8,PP]}} + \ln(\underbrace{0.8}_{\text{np} \rightarrow \text{np pp}}) = -5.6 + -0.2 = -5.8$

Span 5

$$C[3,8,\text{vp}] = \max(\tag{4}$$
$$\underbrace{-3.2}_{\text{C[3,5,VP]}} + \underbrace{-3.3}_{\text{C[6,8,PP]}} + \underbrace{-1.6}_{\text{vp} \rightarrow \text{vp pp}} , \tag{5}$$
$$\underbrace{-0.5}_{\text{C[3,3,V]}} + \underbrace{-5.8}_{\text{C[4,8,NP]}} + \underbrace{-.36}_{\text{vp} \rightarrow \text{v np}}) \tag{6}$$
$$= \max(-8.1, -6.7) = -6.7 \tag{7}$$

Span 7

1. $C[1,8,\mathrm{s}] = \underbrace{-3.9}_{\text{C[1,2,NP]}} + \underbrace{-6.7}_{\text{C[3,8,VP]}} = -10.6$