

Slow, sloppy and expensive:

redundancy and probability in grammar induction

Data Oriented Parsing
Alignment Based Learning
& Exemplar Negotiation

Data Oriented Parsing

- Rens Bod (1991, 1998 “*Beyond Grammar: An Experience-Based Theory of Language*”)

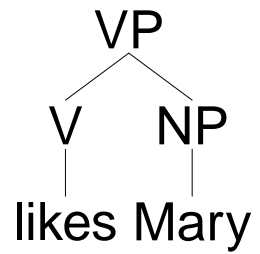
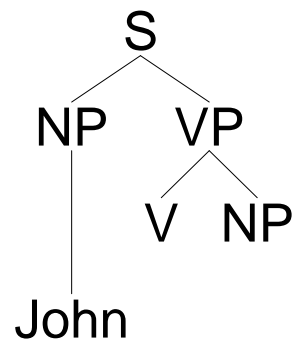
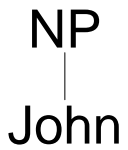
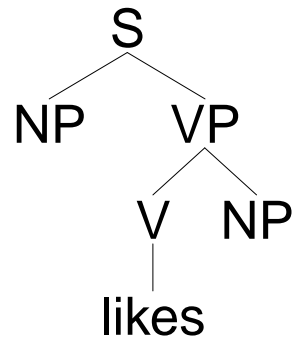
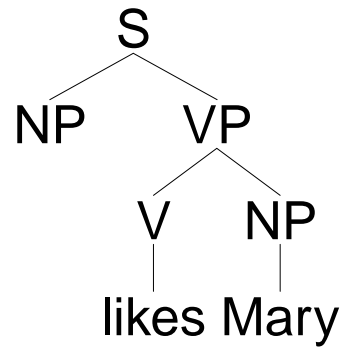
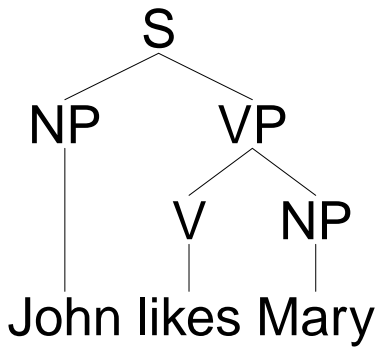
- Key points:

Corpus-based: real life sentences, annotated with phrase-structure (e.g. ATIS)

Massively redundant, unsophisticated representation of grammatical knowledge

Probabilistic (most probable parse)

Formalism



Most Probable Parse

- Composition $A \cdot B$: substitute leftmost nonterminal frontier node in A with B
- Goal: find most probable parse (not most probable derivation)

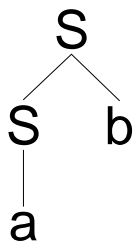
$$\text{a subtree } t : P(t) = \frac{|t|}{|\text{root}(t)|}$$

$$\text{a derivation } d : P(t_1 \cdot \dots \cdot t_n) = \prod_{i=1}^n P(t_i)$$

$$\text{a parse } T : P(T) = \sum_{d:d \rightarrow T} P(d)$$

Weak/strong equivalence

- Weak equivalence: formalisms can generate the same string sets YES
- Strong equivalence: formalisms can generate the same tree sets NO



Tree substitution grammar

$$S \mapsto ab$$

Context-free grammar

$$S \mapsto S b$$

$$S \mapsto a$$

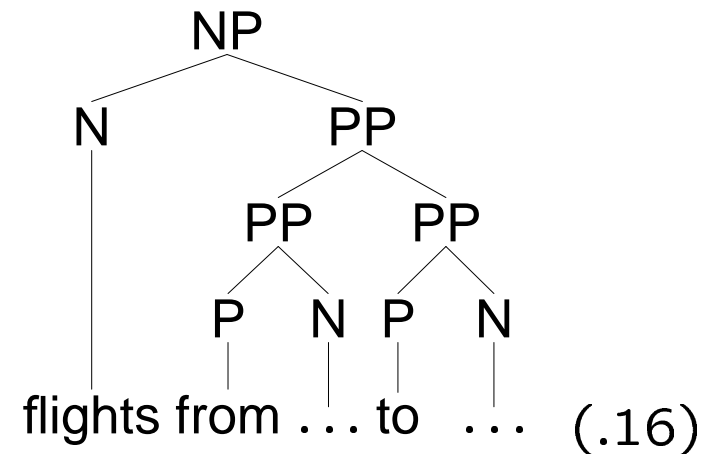
Context-free grammar

Deterministic/stochastic equivalence

- Stochastic equivalence: formalisms can generate the same probability distribution over string or tree setsNO

$PP \mapsto PP\ PP \quad (.2)$
 $PP \mapsto P\ N \quad (.8)$
 $P \mapsto to \quad (.5)$
 $P \mapsto from \quad (.5)$

vs.



Evaluation

DOP works very well: 77% - 85% parse accuracy on test set

This success is due to the difference with CFG's:

- overlapping trees
- large size trees
- lexicalization
- low-frequency trees
- non-head words
- head words

Alignment Based Learning

- Menno van Zaanen (2000, 2001)

- Key Points:

Corpus-based: real-world sentences (e.g. ATIS, OVIS)

Unsupervised: tagging only used for evaluation

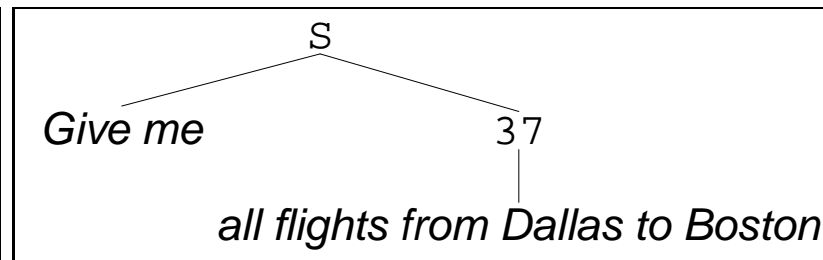
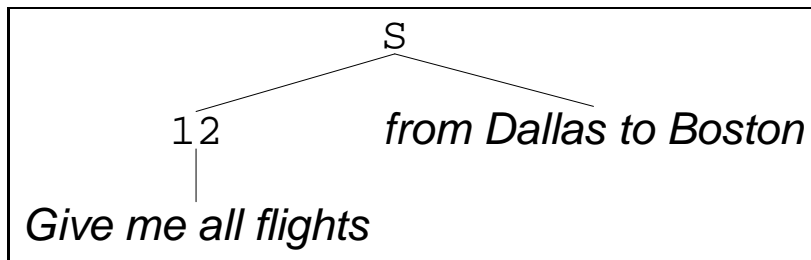
Substitutability *“If our informant accepts DA’F as a repetition of DEF, and if we are similarly able to obtain E’BC as equivalent to ABC, then we say that A and E are mutually substitutable” (Zellig Harris, 1951)*

Alignment Learning Phase

1. a [*Oscar sees Bert*]₁ : a [*Oscar sees [Bert]*]₂₁
b [*Oscar sees Big Bird*]₁ b [*Oscar sees [Big Bird]*]₂₁
2. a [*Oscar sees [Bert]*]₂₁ : a [*Oscar sees [Bert]*]₂₁
b [*Oscar sees Big Bird*]₁ b [*Oscar sees [Big Bird]*]₂₁
3. a [*Oscar sees [Bert]*]₂₁ : a [*Oscar sees [Bert]*]₂₁
b [*Oscar sees [Big Bird]*]₃₁ b [*Oscar sees [Big Bird]*]₂₁
+ 2=3 throughout the hypothesis space

Overlapping hypotheses

- [*Book Delta 128*]₁₂*from Dallas to Boston*
- [*Give me all flights*]₁₂*from Dallas to Boston*
- *Give me* [*all flights from Dallas to Boston*]₃₇
- *Give me* [*help on classes*]₃₇



Selection Learning Phase

- “Alignment learning” generates many hypotheses
- “Selection learning” generates a *treebank* with the most probable set of hypotheses:

$$\text{a hypothesis} : P(h) = \frac{|h|}{|\text{root}(h)|}$$

$$\text{a set of hypotheses} : P(h_1, \dots, h_n) = \sqrt[n]{\prod_{i=1}^n P(h_i)}$$

$$\text{or, equivalently} : \log(P(h_1, \dots, h_n)) = \frac{1}{n} \sum_{i=1}^n \log(P(h_i))$$

Integrating DOP & ABL

The probability of an analysis is the product of the probabilities of its constituent parts, i.e. the probabilities that the used 'building blocks' are indeed the productive building blocks of a language.

The probability of a constituent is given by its frequency in the corpus x the probability that it is part of the maximum-likely hypothesis set.

Incremental Learning

- With each sample sentence, possible constituents should be generated;
- At each step, the expected deviation from the ideal probability assignments should be minimal;
- In the limit, the expected probabilities should be equal to the ideal probabilities.

Incremental Learning

john	john	sees	mary	john	john	sees	silly	pete
sees	N	V	N	sees	N	V	Adj	
mary	?	VP		silly	x	x	NP	
	S			pete	S	VP		N