

Master in Artificial Intelligence

Introduction to Human Language Technologies

8. Syntactic parsing: grammars

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat d'Informàtica de Barcelona



Outline

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

- 1 Syntactic parsing
 - Goal and motivation
 - Types of syntactic structures
- 2 Context Free Grammars (CFGs)
- 3 Probabilistic Context Free Grammars (PCFGs)

Outline

Syntactic
parsing

Goal and motivation

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

- 1 Syntactic parsing
 - Goal and motivation
 - Types of syntactic structures
- 2 Context Free Grammars (CFGs)
- 3 Probabilistic Context Free Grammars (PCFGs)

Goal and motivation

- Syntax studies the combination of words in a sentence.
- Syntactic parsing provides information of the combination of words in a sentence (the syntactic structure).
- Syntactic information is relevant for many NLP applications:
 - Authorship recognition
 - Grammar checking
Ex: 3th-Singular-noun + basic-verb \implies error
 - Machine Translation
Ex: [es] NN+JJ \implies [en] JJ+NN
 - Information Extraction
Ex: $X - [subj] \rightarrow \text{visited} \leftarrow [obj] - Y \implies \text{visit}(X,Y)$
 - ...
- **Goal:** find the syntactic structure associated to a sentence.

Outline

Syntactic
parsing

Types of syntactic
structures

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

1 Syntactic parsing

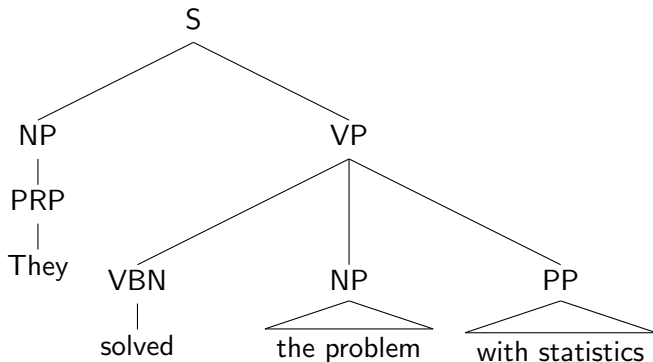
- Goal and motivation

- Types of syntactic structures

2 Context Free Grammars (CFGs)

3 Probabilistic Context Free Grammars (PCFGs)

Constituent tree



Phrase chunking may be seen as the flattening of this structure

[NP They/PRP][VP solved/VBN] [NP the/DT problem/NN] with/IN [NP statistics/NNS]

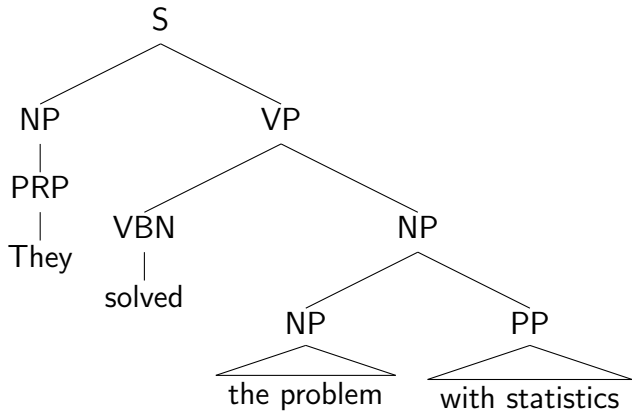
Syntactic
parsing

Types of syntactic
structures

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

Another constituent Tree



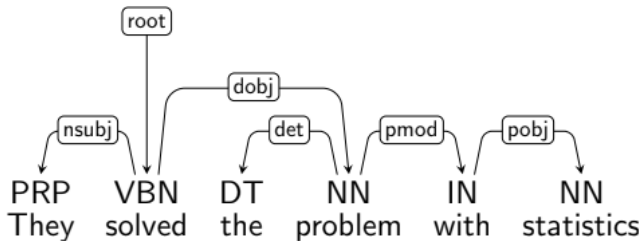
Syntactic
parsing

Types of syntactic
structures

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

Dependency tree



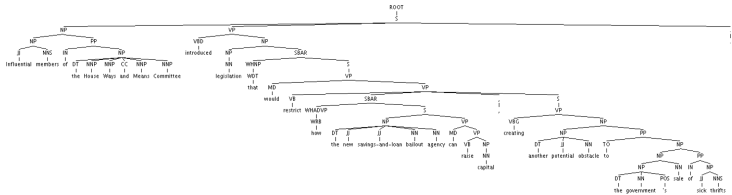
Syntactic
parsing

Types of syntactic
structures

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

A real sentence



Influential members of the House Ways and Means Committee
introduced legislation that would restrict how the new
savings-and-loan bailout agency can raise capital, creating another
potential obstacle to the government's sale of sick thrifts.

Syntactic
parsing

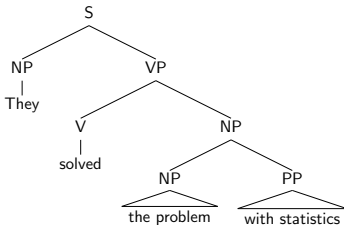
Types of syntactic
structures

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

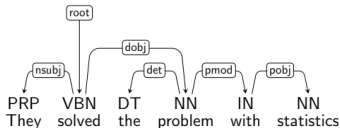
Theories of Syntactic Structure

Constituent Trees



- Main element: constituents (or phrases, or bracketings)
- Constituents = abstract linguistic units
- Results in nested trees

Dependency Trees



- Main element: dependency
- Focus on relations between words
- Handles *free word order* nicely.

Syntactic
parsing

Types of syntactic
structures

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

Outline

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

- 1 Syntactic parsing
 - Goal and motivation
 - Types of syntactic structures
- 2 Context Free Grammars (CFGs)
- 3 Probabilistic Context Free Grammars (PCFGs)

Definition

[Hopcroft and Ullman 1979]

A context free grammar $G = (N, \Sigma, R, S)$ where:

- N is a set of non-terminal symbols
- Σ is a set of terminal symbols
- R is a set of rules of the form $X \rightarrow Y_1 Y_2 \dots Y_n$
for $n \geq 0, X \in N, Y_i \in (N \cup \Sigma)$
- $S \in N$ is a distinguished start symbol

Context Free Grammars, Example

$N = \{S, NP, VP, PP, DT, Vi, Vt, NN, IN\}$

$S = S$

$\Sigma = \{\text{sleeps, saw, man, woman, telescope, the, with, in}\}$

$R =$

S	\Rightarrow	NP	VP
VP	\Rightarrow	Vi	
VP	\Rightarrow	Vt	NP
VP	\Rightarrow	VP	PP
NP	\Rightarrow	DT	NN
NP	\Rightarrow	NP	PP
PP	\Rightarrow	IN	NP

Vi	\Rightarrow	sleeps
Vt	\Rightarrow	saw
NN	\Rightarrow	man
NN	\Rightarrow	woman
NN	\Rightarrow	telescope
DT	\Rightarrow	the
IN	\Rightarrow	with
IN	\Rightarrow	in

Note: S=sentence, VP=verb phrase, NP=noun phrase, PP=prepositional phrase, DT=determiner, Vi=intransitive verb, Vt=transitive verb, NN=noun, IN=preposition

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

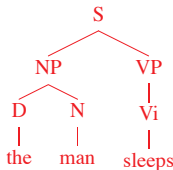
Left-most Derivations in CFGs

A left-most derivation is a sequence of strings $s_1 \dots s_n$, where

- $s_1 = S$, the start symbol
- $s_n \in \Sigma^*$, i.e. s_n is made up of terminal symbols only
- Each s_i for $i = 2 \dots n$ is derived from s_{i-1} by picking the left-most non-terminal X in s_{i-1} and replacing it by some β where $X \rightarrow \beta$ is a rule in R

For example: $[S]$, $[NP VP]$, $[D N VP]$, $[the N VP]$, $[the man VP]$, $[the man Vi]$, $[the man sleeps]$

Representation of a derivation as a tree:



Properties of CFGs

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

- A CFG defines a set of possible derivations
- A string $s \in \Sigma^*$ is in the *language* defined by the CFG if there is at least one derivation which yields s
- Each string in the language generated by the CFG may have more than one derivation (“ambiguity”)

Ambiguities

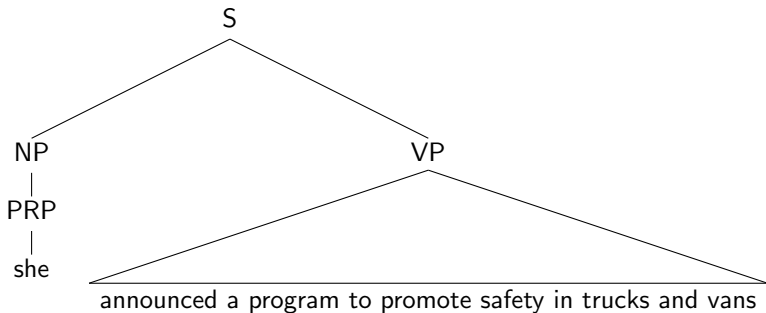
Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

- I cleaned the dishes from dinner
- I cleaned the dishes with detergent
- I cleaned the dishes in my pajamas
- I cleaned the dishes in the sink

Exercise



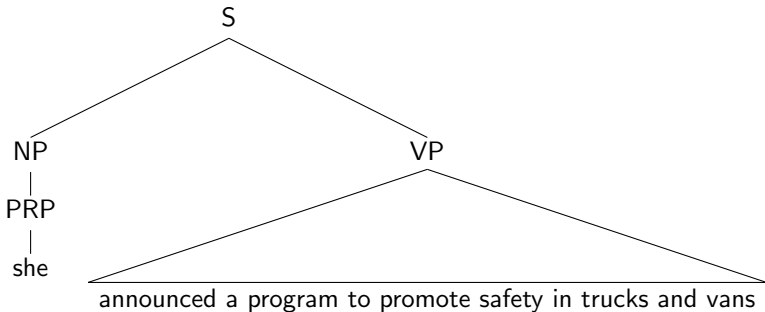
- Give 3 possible interpretations in the form of parse trees
- Provide a CFG to get at least one of the interpretations

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

Exercise



- Give 3 possible interpretations in the form of parse trees
- Provide a CFG to get at least one of the possible parse trees

Probabilistic CFGs can be used to know how likely are the parse trees

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

Outline

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

- 1 Syntactic parsing
 - Goal and motivation
 - Types of syntactic structures
- 2 Context Free Grammars (CFGs)
- 3 Probabilistic Context Free Grammars (PCFGs)

Example

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

S	⇒	NP	VP	1.0
VP	⇒	Vi		0.4
VP	⇒	Vt	NP	0.4
VP	⇒	VP	PP	0.2
NP	⇒	DT	NN	0.3
NP	⇒	NP	PP	0.7
PP	⇒	P	NP	1.0

Vi	⇒	sleeps	1.0
Vt	⇒	saw	1.0
NN	⇒	man	0.7
NN	⇒	woman	0.2
NN	⇒	telescope	0.1
DT	⇒	the	1.0
IN	⇒	with	0.5
IN	⇒	in	0.5

- Probability of a tree t with rules

$$\alpha_1 \rightarrow \beta_1, \alpha_2 \rightarrow \beta_2, \dots, \alpha_n \rightarrow \beta_n$$

is

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

where $q(\alpha \rightarrow \beta)$ is the probability for rule $\alpha \rightarrow \beta$.

Definition

1. A context-free grammar $G = (N, \Sigma, S, R)$.

2. A parameter

$$q(\alpha \rightarrow \beta)$$

for each rule $\alpha \rightarrow \beta \in R$. The parameter $q(\alpha \rightarrow \beta)$ can be interpreted as the conditional probability of choosing rule $\alpha \rightarrow \beta$ in a left-most derivation, given that the non-terminal being expanded is α . For any $X \in N$, we have the constraint

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

In addition we have $q(\alpha \rightarrow \beta) \geq 0$ for any $\alpha \rightarrow \beta \in R$.

Given a parse-tree $t \in \mathcal{T}_G$ containing rules $\alpha_1 \rightarrow \beta_1, \alpha_2 \rightarrow \beta_2, \dots, \alpha_n \rightarrow \beta_n$, the probability of t under the PCFG is

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

Properties of PCFGs

Syntactic
parsing

Context Free
Grammars
(CFGs)

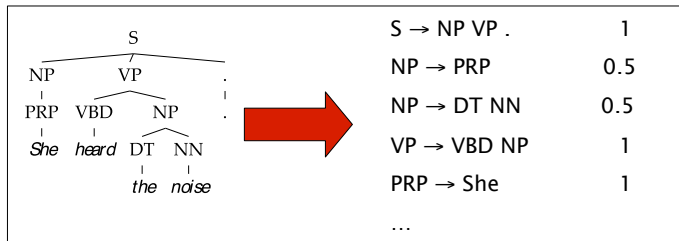
Probabilistic
Context Free
Grammars
(PCFGs)

- Assigns a probability to each *left-most derivation*, or parse-tree, allowed by the underlying CFG
- Say we have a sentence s , set of derivations for that sentence is $\mathcal{T}(s)$. Then a PCFG assigns a probability $p(t)$ to each member of $\mathcal{T}(s)$. i.e., *we now have a ranking in order of probability*.
- The most likely parse tree for a sentence s is

$$\arg \max_{t \in \mathcal{T}(s)} p(t)$$

Learning Treebank Grammars

- Read the grammar rules from a treebank



- Set rule weights by maximum likelihood
- Other approaches are out of this course: PCFG with parent annotations, lexicalized PCFG, PCFG with latent variables

Syntactic
parsing

Context Free
Grammars
(CFGs)

Probabilistic
Context Free
Grammars
(PCFGs)

Maximum Likelihood Estimates

- Algorithm

- 1 Given a treebank, define a CFG **by taking all rules seen in the treebank**
- 2 Maximum Likelihood estimates

$$q(\alpha \rightarrow \beta) = \frac{\text{Count}(\alpha \rightarrow \beta)}{\text{Count}(\alpha)}$$

where the counts are taken from the examples in the treebank.

- Smoothing issues apply here
- Having the appropriate CFG is critical to success

Exercise

Using the following PCFG

$S \rightarrow NP VP$	1.0
$NP \rightarrow NP PP$	0.4
$PP \rightarrow P NP$	1.0
$VP \rightarrow V NP$	0.7
$VP \rightarrow VP PP$	0.3
$P \rightarrow with$	1.0
$V \rightarrow saw$	1.0
$NP \rightarrow astronomers$	0.1
$NP \rightarrow ears$	0.18
$NP \rightarrow saw$	0.04
$NP \rightarrow stars$	0.18
$NP \rightarrow telescope$	0.1

Work with the sentence: '*astronomers saw stars with ears*'

- How many correct parses are there for this sentence?
- Write them along with their probabilities.