Natural Language Processing

Syntactic parsing

Yulia Tsvetkov

yuliats@cs.washington.edu
Announcements

- HW2 is due Monday
  - Note that TAs are not required to provide fast responses over weekends
  - Use TAs office hours this week
  - No extensions beyond “standard” late days due to Thanksgiving
Ambiguity

- I saw a girl with a telescope
Syntactic Parsing

- **INPUT:**
  - The move followed a round of similar increases by other lenders, reflecting a continuing decline in that market

- **OUTPUT:**

A Supervised ML Problem

- Data for parsing experiments:
  - Penn WSJ Treebank = 50,000 sentences with associated trees
  - Usual set-up: 40,000 training, 2,400 test

Canadian Utilities had 1988 revenue of $1.16 billion, mainly from its natural gas and electric utility businesses in Alberta, where the company serves about 800,000 customers. [from Michael Collins slides]
Syntax
Syntax

- The study of the patterns of formation of sentences and phrases from words

  - my dog
    - Pron N
  - the dog
    - Det N
  - the cat
    - Det N
  - and
    - Conj
  - the large cat
    - Det Adj N
  - the black cat
    - Det Adj N
  - ate a sausage
    - V Det N
Parsing

- The process of predicting **syntactic representations**
- Different types of syntactic representations are possible, for example:

```
S
 / \ 
NP  VP
 /  /
PN  N  V  NP
   /    /
  My dog  ate D N
   /  /
  a  sausage
```

**Constituent (a.k.a. phrase-structure) tree**
Constituent trees

- Internal nodes correspond to phrases
  - S – a sentence
  - NP – Noun Phrase: My dog, a sandwich, lakes,..
  - VP – Verb Phrase: ate a sausage, barked, …
  - PP – Prepositional phrases: with a friend, in a car, …

- Nodes immediately above words are PoS tags (aka preterminals)
  - PN – pronoun
  - D – determiner
  - V – verb
  - N – noun
  - P – preposition
Bracketing notation

- It is often convenient to represent a tree as a bracketed sequence

```
(S
  (NP (PN My) (N dog))
  (VP (V ate)
    (NP (D a) (N sausage))
  )
)
```
Parsing

- The process of predicting syntactic representations
- Different types of syntactic representations are possible, for example:

Constituent (a.k.a. phrase-structure) tree

\[ S \]
\[ NP \]
\[ PN \]
\[ My \]
\[ V \]
\[ ate \]
\[ D \]
\[ sausage \]

Dependency tree

\[ root \]
\[ My \]
\[ PN \]
\[ dog \]
\[ N \]
\[ ate \]
\[ V \]
\[ a \]
\[ sausage \]
\[ N \]
Dependency trees

- Nodes are **words** (along with part-of-speech tags)
- Directed arcs encode **syntactic dependencies** between them
- Labels are types of relations between the words
  - **poss** – possessive
  - **dobj** – direct object
  - **nsub** - subject
  - **det** - determiner
Recovering shallow semantics

- Some semantic information can be (approximately) derived from syntactic information
  - Subjects (nsubj) are (often) agents ("initiator / doers for an action")
  - Direct objects (dobj) are (often) patients ("affected entities")
Recovering shallow semantics

- Some semantic information can be (approximately) derived from syntactic information
  - Subjects (nsubj) are (often) agents ("initiator / doers for an action")
  - Direct objects (dobj) are (often) patients ("affected entities")
- But even for agents and patients consider:
  - Mary is baking a cake in the oven
  - A cake is baking in the oven
- In general it is not trivial even for the most shallow forms of semantics
  - E.g., consider prepositions: *in* can encode direction, position, temporal information, …
Constituent and dependency representations

- Constituent trees can (potentially) be converted to dependency trees

- Dependency trees can (potentially) be converted to constituent trees
Constituent trees

- Internal nodes correspond to phrases
  - S – a sentence
  - NP (Noun Phrase): My dog, a sandwich, lakes,..
  - VP (Verb Phrase): ate a sausage, barked, …
  - PP (Prepositional phrases): with a friend, in a car, …

- Nodes immediately above words are PoS tags (aka preterminals)
  - PN – pronoun
  - D – determiner
  - V – verb
  - N – noun
  - P – preposition
Constituency Tests

- How do we know what nodes go in the tree?

- Classic constituency tests:
  - Replacement
  - Movement
    - Passive
    - Clefting
    - Preposing
  - Substitution by proform
  - Modification
  - Coordination/Conjunction
  - Ellipsis/Deletion
Morphology/Syntax/Semantics

- **Syntax:** The study of the patterns of formation of sentences and phrases from word
  - Borders with *semantics* and *morphology* sometimes blurred

"Afyonkarahisarlılaştırılabildiklerimizdenmişsinizcesinee"

in Turkish means "as if you are one of the people that we thought to be originating from Afyonkarahisar" [wikipedia]
English grammar

Product Details (from Amazon)
Hardcover: 1779 pages
Publisher: Longman; 2nd Revised edition
Language: English
ISBN-10: 0582517346
Product Dimensions: 8.4 x 2.4 x 10 inches
Shipping Weight: 4.6 pounds
Context Free Grammar (CFG)
Context Free Grammar (CFG)

Grammar (CFG)

ROOT → S
S → NP VP
NP → DT NN
NP → NN NNS
NP → NP PP
VP → VBP NP
VP → VBP NP PP
PP → IN NP

Lexicon

NN → interest
NNS → raises
VBP → interest
VBZ → raises
...

Other grammar formalisms: LFG, HPSG, TAG, CCG…
CFGs

\[ S \rightarrow NP \ VF \]

\[ VP \rightarrow V \]
\[ VP \rightarrow V \ NP \]
\[ VP \rightarrow VP \ PF \]

\[ NP \rightarrow NP \ PF \]
\[ NP \rightarrow D \ N \]
\[ NP \rightarrow PN \]

\[ PP \rightarrow P \ NF \]

\[ N \rightarrow girl \]
\[ N \rightarrow telescope \]
\[ N \rightarrow sandwich \]
\[ PN \rightarrow I \]
\[ V \rightarrow saw \]
\[ V \rightarrow ate \]
\[ P \rightarrow with \]
\[ P \rightarrow in \]
\[ D \rightarrow a \]
\[ D \rightarrow the \]
CFGs

\[
\begin{align*}
S & \rightarrow NP \ VF \\
NP & \rightarrow D \ N \\
NP & \rightarrow PN \\
VP & \rightarrow V \\
VP & \rightarrow VP NF \\
VP & \rightarrow V NF \\
NP & \rightarrow NP NF \\
PP & \rightarrow P NF
\end{align*}
\]

\[
\begin{align*}
N & \rightarrow \text{girl} \\
N & \rightarrow \text{telescope} \\
N & \rightarrow \text{sandwich} \\
PN & \rightarrow I \\
V & \rightarrow \text{saw} \\
V & \rightarrow \text{ate} \\
P & \rightarrow \text{with} \\
P & \rightarrow \text{in} \\
D & \rightarrow a \\
D & \rightarrow the
\end{align*}
\]
CFGs

\[ S \rightarrow NP \ VF \]
\[ VP \rightarrow V \]
\[ VP \rightarrow VP \ PF \]
\[ NP \rightarrow NP \ PF \]
\[ NP \rightarrow D \ N \]
\[ NP \rightarrow PN \]
\[ PP \rightarrow P \ NF \]

\[ N \rightarrow girl \]
\[ N \rightarrow telescope \]
\[ N \rightarrow sandwich \]
\[ PN \rightarrow I \]
\[ V \rightarrow saw \]
\[ V \rightarrow ate \]
\[ P \rightarrow with \]
\[ P \rightarrow in \]
\[ D \rightarrow a \]
\[ D \rightarrow the \]
CFGs

```
S → NP VP
VP → V
VP → V NP
VP → VP PF
NP → NP PF
NP → D N
NP → PN
PP → P NF
```

```
N → girl
N → telescope
N → sandwich
PN → I
V → saw
V → ate
P → with
P → in
D → a
D → the
```
CFGs

\[
S \rightarrow NP \ VF \\
VP \rightarrow V \\
VP \rightarrow VP \ PF \\
NP \rightarrow NP \ PF \\
NP \rightarrow D \ N \\
NP \rightarrow PN \\
PP \rightarrow P \ NF \\
N \rightarrow girl \\
N \rightarrow telescope \\
N \rightarrow sandwich \\
PN \rightarrow I \\
V \rightarrow saw \\
V \rightarrow ate \\
P \rightarrow with \\
P \rightarrow in \\
D \rightarrow a \\
D \rightarrow the
\]
CFGs

\[
S \rightarrow NP \ VP \\
VP \rightarrow V \\
VP \rightarrow V \ NP \\
VP \rightarrow VP \ PP \\
NP \rightarrow NP \ PP \\
NP \rightarrow D \ N \\
NP \rightarrow PN \\
PP \rightarrow P \ NF \\
V \rightarrow saw \\
N \rightarrow girl \\
N \rightarrow telescope \\
N \rightarrow sandwich \\
PN \rightarrow I \\
P \rightarrow with \\
D \rightarrow a \\
D \rightarrow the
\]
CFGs

\[
S \rightarrow NP \ VF
\]

\[
N \rightarrow \text{girl}
\]

\[
N \rightarrow \text{telescope}
\]

\[
N \rightarrow \text{sandwich}
\]

\[
VP \rightarrow V
\]

\[
VP \rightarrow V \ NF
\]

\[
VP \rightarrow VP \ PF
\]

\[
V \rightarrow \text{saw}
\]

\[
V \rightarrow \text{ate}
\]

\[
PN \rightarrow I
\]

\[
P \rightarrow \text{with}
\]

\[
PP \rightarrow P \ NF
\]

\[
P \rightarrow \text{in}
\]

\[
D \rightarrow a
\]

\[
D \rightarrow \text{the}
\]

\[
NP \rightarrow NP \ PF
\]

\[
NP \rightarrow D \ N
\]

\[
NP \rightarrow PN
\]

\[
I \rightarrow \text{saw}
\]
CFGs

S → NP VP

NP → PN I

VP → V NP

NP → D N

PP → P NF

N → girl

V → saw

N → telescope

VP → VP PP

N → sandwich

PN → I

V → ate

P → with

P → in

D → a

D → the
Treebank Sentences

(S (NP-SBJ The move)
 (VP followed
  (NP (NP a round)
   (PP of
    (NP (NP similar increases)
     (PP by
      (NP other lenders))
     (PP against
      (NP Arizona real estate loans))))))

(S-ADV (NP-SBJ *)
 (VP reflecting
  (NP (NP a continuing decline)
   (PP-LOC in
    (NP that market))))))
Context-Free Grammars

● A context-free grammar is a 4-tuple \( <N, T, S, R> \)
  ○ \( N \) : the set of non-terminals
    ■ **Phrasal categories**: S, NP, VP, ADJP, etc.
    ■ **Parts-of-speech** (pre-terminals): NN, JJ, DT, VB
  ○ \( T \) : the set of terminals (the words)
  ○ \( S \) : the start symbol
    ■ Often written as ROOT or TOP
    ■ Not usually the sentence non-terminal S
  ○ \( R \) : the set of rules
    ■ Of the form \( X \rightarrow Y_1 Y_2 \ldots Y_k \), with \( X, Y_i \in N \)
    ■ Examples: \( S \rightarrow NP \ VP \), \( VP \rightarrow VP \ CC \ VP \)
    ■ Also called rewrites, productions, or local trees
An example grammar

\[ N = \{S, VP, NP, PP, N, V, PN, P\} \]

\[ T = \{girl, telescope, sandwich, I, saw, ate, with, in, a, the\} \]

\[ S = \{S\} \]

\[ R \]

\[ S \rightarrow NP \ VF \]

\[ VP \rightarrow V \]

\[ VP \rightarrow V \ NF \]

\[ VP \rightarrow VP \ PF \]

\[ NP \rightarrow NP \ PF \]

\[ NP \rightarrow D \ N \]

\[ NP \rightarrow PN \]

\[ PP \rightarrow P \ NF \]

Preterminal rules

\[ N \rightarrow girl \]

\[ N \rightarrow telescope \]

\[ N \rightarrow sandwich \]

\[ PN \rightarrow I \]

\[ V \rightarrow saw \]

\[ V \rightarrow ate \]

\[ P \rightarrow with \]

\[ P \rightarrow in \]

\[ D \rightarrow a \]

\[ D \rightarrow the \]
Why context-free?

What can be a sub-tree is only affected by what the phrase type is (VP) but not the context.
Why context-free?

What can be a sub-tree is only affected by what the phrase type is (VP) but not the context.

Not grammatical.
Ambiguities
Coordination ambiguity

- Here, the coarse VP and NP categories cannot enforce subject-verb agreement in number resulting in the coordination ambiguity

"Bark" can refer both to a noun or a verb

This tree would be ruled out if the context would be somehow captured (subject-verb agreement)
Why parsing is hard?  Ambiguity

- Prepositional phrase attachment ambiguity
PP Ambiguity

*Put the block in the box on the table in the kitchen*

3 prepositional phrases, 5 interpretations:

- Put the block ((in the box on the table) in the kitchen)
- Put the block (in the box (on the table in the kitchen))
- Put ((the block in the box) on the table) in the kitchen.
- Put (the block (in the box on the table)) in the kitchen.
- Put (the block in the box) (on the table in the kitchen)
**PP Ambiguity**

*Put the block in the box on the table in the kitchen*

3 prepositional phrases, 5 interpretations:

- Put the block ((in the box on the table) in the kitchen)
- Put the block (in the box (on the table in the kitchen))
- ...

A general case:

- (((())))  (())()  ()()()  (())()  (()())

\[ Cat_n = \binom{2n}{n} - \binom{2n}{n-1} \sim \frac{4^n}{n^{3/2}\sqrt{\pi}} \]

1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, ...
Canadian Utilities had 1988 revenue of $1.16 billion, mainly from its natural gas and electric utility businesses in Alberta, where the company serves about 800,000 customers.

[from Michael Collins slides]
Syntactic Ambiguities I

- **Prepositional phrases:**
  - They cooked the beans in the pot on the stove with handles.

- **Particle vs. preposition:**
  - The puppy tore up the staircase.

- **Complement structures**
  - The tourists objected to the guide that they couldn’t hear.
  - She knows you like the back of her hand.

- **Gerund vs. participial adjective**
  - Visiting relatives can be boring.
  - Changing schedules frequently confused passengers.
Syntactic Ambiguities II

- **Modifier scope within NPs**
  - impractical design requirements
    - plastic cup holder

- **Multiple gap constructions**
  - The chicken is ready to eat.
  - The contractors are rich enough to sue.

- **Coordination scope:**
  - Small rats and mice can squeeze into holes or cracks in the wall.
How to Deal with Ambiguity?

- We want to score all the derivations to encode how plausible they are.

*Put the block in the box on the table in the kitchen*
Probabilistic Context Free Grammar (PCFG)
Probabilistic Context-Free Grammars

- A context-free grammar is a 4-tuple \(<N, T, S, R>\)
  - \(N\) : the set of non-terminals
    - **Phrasal categories**: S, NP, VP, ADJP, etc.
    - **Parts-of-speech** (pre-terminals): NN, JJ, DT, VB
  - \(T\) : the set of terminals (the words)
  - \(S\) : the start symbol
    - Often written as ROOT or TOP
    - Not usually the sentence non-terminal S
  - \(R\) : the set of rules
    - Of the form \(X \rightarrow Y_1 Y_2 \ldots Y_k\), with \(X, Y_i \in N\)
    - Examples: \(S \rightarrow NP \ VP\), \(VP \rightarrow VP \ CC \ VP\)
    - Also called rewrites, productions, or local trees

- A PCFG adds:
  - A top-down **production probability** per rule \(P(Y_1 Y_2 \ldots Y_k | X)\)
PCFGs

Associate probabilities with the rules:

\[ p(X \rightarrow \alpha) \]

\[ \forall X \rightarrow \alpha \in R: \quad 0 \leq p(X \rightarrow \alpha) \leq 1 \]

\[ \forall X \in N: \quad \sum_{\alpha: X \rightarrow \alpha \in R} p(X \rightarrow \alpha) = 1 \]

\[ S \rightarrow NP \ VF \] 1.0  

\[ (NP \ A \ girl) \ (VP \ ate \ a \ sandwich) \]

\[ VP \rightarrow V \] 0.2 

\[ (VP \ ate) \ (NP \ a \ sandwich) \]

\[ VP \rightarrow VP \ PF \] 0.4  

\[ (VP \ saw \ a \ girl) \ (PP \ with \ ...) \]

\[ NP \rightarrow NP \ PF \] 0.3  

\[ (NP \ a \ girl) \ (PP \ with \ ...) \]

\[ NP \rightarrow D \ N \] 0.5  

\[ (D \ a) \ (N \ sandwich) \]

\[ NP \rightarrow PN \] 0.2 

\[ (P \ with) \ (NP \ with \ a \ sandwich) \]

Now we can score a tree as a product of probabilities corresponding to the used rules.
PCFGs

\[
p(T) = S \rightarrow NP \ VP \ 1.0
\]

\[
VP \rightarrow V \ 0.2
\]

\[
VP \rightarrow VP \ PF \ 0.4
\]

\[
NP \rightarrow NP \ PF \ 0.3
\]

\[
NP \rightarrow D \ N \ 0.5
\]

\[
NP \rightarrow PN \ 0.2
\]

\[
PP \rightarrow P \ NF \ 1.0
\]

\[
N \rightarrow girl \ 0.2
\]

\[
N \rightarrow telescope \ 0.7
\]

\[
N \rightarrow sandwich \ 0.1
\]

\[
PN \rightarrow I \ 1.0
\]

\[
V \rightarrow saw \ 0.5
\]

\[
V \rightarrow ate \ 0.5
\]

\[
P \rightarrow with \ 0.6
\]

\[
P \rightarrow in \ 0.4
\]

\[
D \rightarrow a \ 0.3
\]

\[
D \rightarrow the \ 0.7
\]
PCFGs

\[ S \rightarrow NP \ VF \ 1.0 \]
\[ VP \rightarrow V \ 0.2 \]
\[ VP \rightarrow V \ NF \ 0.4 \]
\[ VP \rightarrow VP \ PF \ 0.4 \]
\[ NP \rightarrow NP \ PF \ 0.3 \]
\[ NP \rightarrow D \ N \ 0.5 \]
\[ NP \rightarrow PN \ 0.2 \]
\[ PP \rightarrow P \ NF \ 1.0 \]

\[ N \rightarrow girl \ 0.2 \]
\[ N \rightarrow telescope \ 0.7 \]
\[ N \rightarrow sandwich \ 0.1 \]
\[ PN \rightarrow I \ 1.0 \]
\[ V \rightarrow saw \ 0.5 \]
\[ V \rightarrow ate \ 0.5 \]
\[ P \rightarrow with \ 0.6 \]
\[ P \rightarrow in \ 0.4 \]
\[ D \rightarrow a \ 0.3 \]
\[ D \rightarrow the \ 0.7 \]

\[ p(T) = 1.0 \times \]
PCFGs

\[ p(T) = 1.0 \times 0.2 \times \]

\[
S \rightarrow NP \ VF \ 1.0 \\
VP \rightarrow V \ NF \ 0.2 \\
VP \rightarrow V \ PP \ 0.4 \\
NP \rightarrow NP \ PP \ 0.3 \\
NP \rightarrow D \ N \ 0.5 \\
NP \rightarrow PN \ 0.2 \\
PP \rightarrow P \ NF \ 1.0 \\
N \rightarrow girl \ 0.2 \\
N \rightarrow telescope \ 0.7 \\
N \rightarrow sandwich \ 0.1 \\
PN \rightarrow I \ 1.0 \\
V \rightarrow saw \ 0.5 \\
V \rightarrow ate \ 0.5 \\
P \rightarrow with \ 0.6 \\
P \rightarrow in \ 0.4 \\
D \rightarrow a \ 0.3 \\
D \rightarrow the \ 0.7 \]
PCFGs

\[ p(T) = 1.0 \times 0.2 \times 1.0 \times \]

\[
S \to NP \ VF \ 1.0
\]

\[
VP \to V \ 0.2
\]

\[
VP \to V \ NF \ 0.4
\]

\[
VP \to VP \ PP \ 0.4
\]

\[
NP \to NP \ PP \ 0.3
\]

\[
NP \to D \ N \ 0.5
\]

\[
NP \to PN \ 0.2
\]

\[
PP \to P \ NF \ 1.0
\]

\[
N \to girl \ 0.2
\]

\[
N \to telescope \ 0.7
\]

\[
N \to sandwich \ 0.1
\]

\[
PN \to I \ 1.0
\]

\[
V \to saw \ 0.5
\]

\[
V \to ate \ 0.5
\]

\[
P \to with \ 0.6
\]

\[
P \to in \ 0.4
\]

\[
D \to a \ 0.3
\]

\[
D \to the \ 0.7
\]
PCFGs

\[ S \rightarrow NP \ VP \ 1.0 \]

\[ VP \rightarrow V \ 0.2 \]

\[ VP \rightarrow V \ NP \ 0.4 \]

\[ VP \rightarrow VP \ PF \ 0.4 \]

\[ NP \rightarrow NP \ PF \ 0.3 \]

\[ NP \rightarrow D \ N \ 0.5 \]

\[ NP \rightarrow PN \ 0.2 \]

\[ PP \rightarrow P \ NF \ 1.0 \]

\[ N \rightarrow girl \ 0.2 \]

\[ N \rightarrow telescope \ 0.7 \]

\[ N \rightarrow sandwich \ 0.1 \]

\[ PN \rightarrow I \ 1.0 \]

\[ V \rightarrow saw \ 0.5 \]

\[ V \rightarrow ate \ 0.5 \]

\[ P \rightarrow with \ 0.6 \]

\[ P \rightarrow in \ 0.4 \]

\[ D \rightarrow a \ 0.3 \]

\[ D \rightarrow the \ 0.7 \]

\[ p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times \]
PCFGs

\[ p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times 0.5 \times \]

\[
S \rightarrow NP \ VF \ 1.0 \\
NP \rightarrow I \ 0.2 \\
VP \rightarrow V \ 0.4 \\
NP \rightarrow D \ N \ 0.5 \\
PP \rightarrow P \ NF \ 1.0 \\
PN \rightarrow I \ 1.0 \\
V \rightarrow saw \ 0.5 \\
V \rightarrow ate \ 0.5 \\
P \rightarrow with \ 0.6 \\
P \rightarrow in \ 0.4 \\
D \rightarrow a \ 0.3 \\
D \rightarrow the \ 0.7 \\
N \rightarrow girl \ 0.2 \\
N \rightarrow telescope \ 0.7 \\
N \rightarrow sandwich \ 0.1 \\
\]

Undergrad NLP 2022
PCFGs

\[ p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times 0.5 \times 0.3 \times \]

- \[ S \rightarrow NP \ VP \ 1.0 \]
- \[ N \rightarrow girl \ 0.2 \]
- \[ N \rightarrow telescope \ 0.7 \]
- \[ N \rightarrow sandwich \ 0.1 \]
- \[ PN \rightarrow I \ 1.0 \]
- \[ V \rightarrow saw \ 0.5 \]
- \[ V \rightarrow ate \ 0.5 \]
- \[ P \rightarrow with \ 0.6 \]
- \[ P \rightarrow in \ 0.4 \]
- \[ D \rightarrow a \ 0.3 \]
- \[ D \rightarrow the \ 0.7 \]
PCFGs

\[ p(T) = 1.0 \times 0.2 \times 1.0 \times 0.4 \times 0.5 \times 0.3 \times 0.5 \times 0.3 \times 0.2 \times 1.0 \times 0.6 \times 0.5 \times 0.3 \times 0.7 \times 2.26 \times 10^{-5} \]
PCFG Estimation
ML estimation

- A treebank: a collection of sentences annotated with constituent trees

- An estimated probability of a rule (maximum likelihood estimates)

\[ p(X \rightarrow \alpha) = \frac{C(X \rightarrow \alpha)}{C(X)} \]

- Smoothing is helpful
  - Especially important for preterminal rules

The number of times the rule used in the corpus

The number of times the nonterminal X appears in the treebank
Parsing evaluation
Parsing evaluation

- **Intrinsic** evaluation:
  - **Automatic**: evaluate against annotation provided by human experts (gold standard) according to some predefined measure
  - **Manual**: … according to human judgment

- **Extrinsic** evaluation: score syntactic representation by comparing how well a system using this representation performs on some task
  - E.g., use syntactic representation as input for a semantic analyzer and compare results of the analyzer using syntax predicted by different parsers.
Standard evaluation setting in parsing

- Automatic intrinsic evaluation is used: parsers are evaluated against gold standard by provided by linguists
  - There is a standard split into the parts:
    - training set: used for estimation of model parameters
    - development set: used for tuning the model (initial experiments)
    - test set: final experiments to compare against previous work
Automatic evaluation of constituent parsers

- **Exact match**: percentage of trees predicted correctly
- **Bracket score**: scores how well individual phrases (and their boundaries) are identified

The most standard measure; we will focus on it
Brackets scores

- The most standard score is bracket score.
- It regards a tree as a collection of brackets.
- The set of brackets predicted by a parser is compared against the set of brackets in the tree annotated by a linguist.
- Precision, recall and F1 are used as scores.

Subtree signatures for CKY
Preview: F1 bracket score

- Treebank PCFG
- Unlexicalized PCFG (Klein and Manning, 2003)
- Lexicalized PCFG (Collins, 1999)
- Automatically Induced PCFG (Petrov et al., 2006)
- The best results reported (as of 2012)
CKY Parsing
Parsing

- **Parsing is search** through the space of all possible parses
  - e.g., we may want either any parse, all parses or the highest scoring parse (if PCFG):
    \[
    \text{arg max } P(T) \\
    T \in G(x)
    \]

- **Bottom-up:**
  - One starts from words and attempt to construct the full tree

- **Top-down**
  - Start from the start symbol and attempt to expand to get the sentence
CKY algorithm (aka CYK)

- **Cocke-Kasami-Younger** algorithm
  - Independently discovered in late 60s / early 70s

- An efficient bottom up parsing algorithm for (P)CFGs
  - can be used both for the recognition and parsing problems
  - Very important in NLP (and beyond)

- We will start with the non-probabilistic version
Constraints on the grammar

- The basic CKY algorithm supports only rules in the Chomsky Normal Form (CNF):

\[
\begin{align*}
C & \rightarrow x \\
C & \rightarrow C_1 \, C_2
\end{align*}
\]

Unary preterminal rules (generation of words given PoS tags):

\[
N \rightarrow \text{telescope} \quad D \rightarrow \text{the}
\]

Binary inner rules:

\[
S \rightarrow NP \, VF \quad NP \rightarrow D \, N
\]
Constraints on the grammar

- The basic CKY algorithm supports only rules in the Chomsky Normal Form (CNF):
  \[ C \rightarrow x \]
  \[ C \rightarrow C_1 C_2 \]

- Any CFG can be converted to an equivalent CNF
  - Equivalent means that they define the same language
  - However (syntactic) trees will look differently
  - It is possible to address it by defining such transformations that allows for easy reverse transformation
Transformation to CNF form

- What one need to do to convert to CNF form
  - Get rid of rules that mix terminals and non-terminals
  - Get rid of unary rules: \[ C \rightarrow C_1 \]
  - Get rid of N-ary rules: \[ C \rightarrow C_1 C_2 \ldots C_n \ (n > 2) \]

Crucial to process them, as required for efficient parsing
Transformation to CNF form: binarization

Consider

\[ NP \rightarrow DT \; NNP \; VBG \; NN \]

- How do we get a set of binary rules which are equivalent?
Transformation to CNF form: binarization

- Consider

\[ NP \rightarrow DT \ NNP \ VBG \ NN \]

\[ \begin{array}{c}
NP \\
\downarrow \\
DT \ \text{the} \\
\downarrow \\
NNP \ \text{Dutch} \\
\downarrow \\
VBG \ \text{publishing} \\
\downarrow \\
NN \ \text{group}
\end{array} \]

- How do we get a set of binary rules which are equivalent?

\[ NP \rightarrow DT \ X \]
\[ X \rightarrow NNP \ Y \]
\[ Y \rightarrow VBG \ NN \]
Transformation to CNF form: binarization

- Consider

\[ NP \rightarrow DT \ NNP \ VBG \ NN \]

\[
\begin{array}{c}
NP \\
\downarrow \\
DT \\
\downarrow \\
\text{the} \\
\downarrow \\
NNP \\
\downarrow \\
\text{Dutch} \\
\downarrow \\
VBG \\
\downarrow \\
\text{publishing} \\
\downarrow \\
NN \\
\end{array}
\]

- How do we get a set of binary rules which are equivalent?

\[ NP \rightarrow DT \ X \]
\[ X \rightarrow NNP \ Y \]
\[ Y \rightarrow VBG \ NN \]

- A more systematic way to refer to new non-terminals

\[ NP \rightarrow DT \ @NP|DT \]
\[ @NP|DT \rightarrow NNP \ @NP|DT.NNP \]
\[ @NP|DT.NNP \rightarrow VBG \ NN \]
Transformation to CNF form: binarization

- Instead of binarizing tuples we can binarize trees on preprocessing:

  Also known as **lossless Markovization** in the context of PCFGs

  Can be easily reversed on postprocessing
CKY: Parsing task

- We are given
  - a grammar \(<N, T, S, R>\)
  - a sequence of words \(w = (w_1, w_2, \ldots, w_n)\)

- Our goal is to produce a parse tree for \(w\)
CKY: Parsing task

- We are given
  - a grammar \( <N, T, S, R> \)
  - a sequence of words \( w = (w_1, w_2, \ldots, w_n) \)
- Our goal is to produce a parse tree for \( w \)
- We need an easy way to refer to substrings of \( w \)

span \((i, j)\) refers to words between fenceposts \(i\) and \(j\)

indices refer to fenceposts
Parsing one word

\[ C \rightarrow w_i \]
Parsing one word

\[ C \rightarrow w_i \]
Parsing one word

$C \rightarrow w_i$

covers all words between $i-1$ and $i$
Parsing longer spans

$C \rightarrow C_1 \ C_2$

Check through all $C_1$, $C_2$, mid

covers all words btw $min$ and $mid$
covers all words btw $mid$ and $max$
Parsing longer spans

\[ C \rightarrow C_1 \quad C_2 \]

Check through all C1, C2, mid

covers all words btw min and mid

covers all words btw mid and max
Parsing longer spans

C

covers all words between min and max
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

max = 1  max = 2  max = 3

min = 0

min = 1

min = 2

Preterminal rules

NP → N
NP → N NP

Inner rules

VP → M V
VP → V

NP → N

N → can
N → lead
N → poison

M → can
M → must

V → poison
V → lead

Chart (aka parsing triangle)
\[ S \rightarrow NP \ VP \]

Inner rules:

\[ NP \rightarrow N \]
\[ NP \rightarrow N \ NP \]

Preterminal rules:

\[ VP \rightarrow M \ V \]
\[ VP \rightarrow V \]

\[ N \rightarrow can \]
\[ N \rightarrow lead \]
\[ N \rightarrow poison \]

\[ M \rightarrow can \]
\[ M \rightarrow must \]

\[ V \rightarrow poison \]
\[ V \rightarrow lead \]
$S \rightarrow NP\ VP$

$VP \rightarrow M\ V$

$VP \rightarrow V$

$NP \rightarrow N$

$NP \rightarrow N\ NP$

$N \rightarrow \text{can}$

$N \rightarrow \text{lead}$

$N \rightarrow \text{poison}$

$M \rightarrow \text{can}$

$M \rightarrow \text{must}$

$V \rightarrow \text{poison}$

$V \rightarrow \text{lead}$
### Preterminal rules

- $S \rightarrow NP \ VP$

### Inner rules

- $VP \rightarrow M \ V$
- $VP \rightarrow V$
- $NP \rightarrow N$
- $NP \rightarrow N \ NP$

### Grammar rules

- $N \rightarrow can$
- $N \rightarrow lead$
- $N \rightarrow poison$
- $M \rightarrow can$
- $M \rightarrow must$
- $V \rightarrow poison$
- $V \rightarrow lead$
**Preterminal rules**

<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\begin{array}{c|c|c|c|c}
 & 1 & 4 & 6 & S' \tabularnewline \hline
1 & & & & \tabularnewline 2 & & & & \tabularnewline 3 & & & & \tabularnewline \end{array}
\]

\[
\begin{align*}
V P & \rightarrow M V \\
V P & \rightarrow V \\
N P & \rightarrow N \\
N P & \rightarrow N N P \\
N & \rightarrow can \\
N & \rightarrow lead \\
N & \rightarrow poison \\
M & \rightarrow can \\
M & \rightarrow must \\
V & \rightarrow poison \\
V & \rightarrow lead
\end{align*}
\]
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

max = 1  max = 2  max = 3

min = 0  min = 1  min = 2

$S \rightarrow NP \ VP$

$VP \rightarrow M \ V$
$VP \rightarrow V$

$NP \rightarrow N$
$NP \rightarrow N \ NP$

$N \rightarrow can$
$N \rightarrow lead$
$N \rightarrow poison$

$M \rightarrow can$
$M \rightarrow must$

$V \rightarrow poison$
$V \rightarrow lead$
Preterminal rules

- $V \rightarrow \text{poison}$
- $V \rightarrow \text{lead}$
- $M \rightarrow \text{must}$
- $N \rightarrow \text{can}$
- $N \rightarrow \text{lead}$
- $N \rightarrow \text{poison}$
- $NP \rightarrow N \ NP$
- $VP \rightarrow M \ V$
- $VP \rightarrow V$

Inner rules

- $0$ lead
- $1$ can
- $2$ poison

min = 2
min = 1
min = 0
max = 1
max = 2
max = 3
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

$S \rightarrow NP \ VP$

$VP \rightarrow M \ V$
$VP \rightarrow V$

$NP \rightarrow N$
$NP \rightarrow N \ NP$

$N \rightarrow can$
$N \rightarrow lead$
$N \rightarrow poison$

$M \rightarrow can$
$M \rightarrow must$

$V \rightarrow poison$
$V \rightarrow lead$
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

max = 1  max = 2  max = 3

min = 0  min = 1  min = 2

1 \ N, V
2 \ N, M
3 \ N, V

\( S \rightarrow NP \ VP \)

\( VP \rightarrow M \ V \)
\( VP \rightarrow V \)

\( NP \rightarrow N \)
\( NP \rightarrow N \ NP \)

\( N \rightarrow can \)
\( N \rightarrow lead \)
\( N \rightarrow poison \)

\( M \rightarrow can \)
\( M \rightarrow must \)

\( V \rightarrow poison \)
\( V \rightarrow lead \)
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\begin{array}{ccc}
\text{max = 1} & \text{max = 2} & \text{max = 3} \\
\hline
1 & N, V & \text{?} \\
\text{min = 0} & NP, VP \\
2 & N, M & NP \\
\text{min = 1} \\
3 & N, V & NP, VP \\
\text{min = 2} \\
\end{array}
\]

\[
S \rightarrow NP \ VP
\]

Inner rules:

\[
VP \rightarrow M \ V \\
VP \rightarrow V
\]

Preterminal rules:

\[
NP \rightarrow N \\
NP \rightarrow N \ NP
\]

\[
N \rightarrow \text{can} \\
N \rightarrow \text{lead} \\
N \rightarrow \text{poison}
\]

\[
M \rightarrow \text{can} \\
M \rightarrow \text{must}
\]

\[
V \rightarrow \text{poison} \\
V \rightarrow \text{lead}
\]
Undergrad NLP 2022

Yulia Tsvetkov

<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

max = 1  max = 2  max = 3

min = 0

min = 1

min = 2

Inner rules

\[
S \rightarrow NP \ VP
\]

\[
VP \rightarrow M \ V \\
VP \rightarrow V
\]

\[
NP \rightarrow N \\
NP \rightarrow N \ NP
\]

Preterminal rules

\[
N \rightarrow can \\
N \rightarrow lead \\
N \rightarrow poison
\]

\[
M \rightarrow can \\
M \rightarrow must
\]

\[
V \rightarrow poison \\
V \rightarrow lead
\]
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
\begin{array}{c|c|c}
\text{max = 1} & \text{max = 2} & \text{max = 3} \\
\hline
1. N, V & 4. NP & \\
NP, VP & \\
\hline
2. N, M & 5. ? & \\
NP & \\
\hline
3. N, V & & \\
NP, VP & \\
\end{array}
\]

Preterminal rules:
- \( S \to NP \ VP \)
- \( VP \to M \ V \)
- \( VP \to V \)
- \( NP \to N \)
- \( NP \to N \ NP \)
- \( N \to \text{can} \)
- \( N \to \text{lead} \)
- \( N \to \text{poison} \)
- \( M \to \text{can} \)
- \( M \to \text{must} \)
- \( V \to \text{poison} \)
- \( V \to \text{lead} \)
Undergrad NLP 2022

Preterminal rules

<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

max = 1  
max = 2  
max = 3

min = 0  
\[ N, V \]
\[ NP, VP \]

min = 1  
\[ N, M \]
\[ NP \]

min = 2  
\[ N, V \]
\[ NP, VP \]

Inner rules

\[ S \rightarrow NP \ V P \]
\[ VP \rightarrow M \ V \]
\[ VP \rightarrow V \]
\[ NP \rightarrow N \]
\[ NP \rightarrow N \ NP \]

\[ N \rightarrow can \]
\[ N \rightarrow lead \]
\[ N \rightarrow poison \]
\[ M \rightarrow can \]
\[ M \rightarrow must \]
\[ V \rightarrow poison \]
\[ V \rightarrow lead \]
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\[
S \rightarrow NP \ VP
\]

\[
VP \rightarrow M \ V \\
VP \rightarrow V
\]

\[
NP \rightarrow N \\
NP \rightarrow N NP
\]

\[
N \rightarrow can \\
N \rightarrow lead \\
N \rightarrow poison
\]

\[
M \rightarrow can \\
M \rightarrow must
\]

\[
V \rightarrow poison \\
V \rightarrow lead
\]
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Preterminal rules:

- $S \rightarrow NP \ VP$
- $VP \rightarrow M \ V$
- $VP \rightarrow V$
- $NP \rightarrow N$
- $NP \rightarrow N \ NP$

Inner rules:

- $N \rightarrow can$
- $N \rightarrow lead$
- $N \rightarrow poison$
- $M \rightarrow can$
- $M \rightarrow must$
- $V \rightarrow poison$
- $V \rightarrow lead$
Undergrad NLP 2022

Yulia Tsvetkov

Preterminal rules

S → NP VP

Inner rules

VP → M V
VP → V

NP → N
NP → N NP

N → can
N → lead
N → poison

M → can
M → must

V → poison
V → lead

<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

mid=1

max = 1  max = 2  max = 3

min = 0

1  N, V
   NP, VP

2  N, M
   NP

3  N, V
   NP, VP

4  NP

5  S, VP, NP

6  S, NP
<table>
<thead>
<tr>
<th>lead</th>
<th>can</th>
<th>poison</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Preterminal rules**

\[
S \rightarrow \text{NP VP}
\]

**Inner rules**

\[
VP \rightarrow M V \\
VP \rightarrow V
\]

\[
NP \rightarrow N \\
NP \rightarrow N NP
\]

**Non-terminals**

<table>
<thead>
<tr>
<th>min = 0</th>
<th>min = 1</th>
<th>min = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (N, V) (NP, VP)</td>
<td>2 (N, M) (NP)</td>
<td>3 (N, V) (NP, VP)</td>
</tr>
<tr>
<td>4 (NP)</td>
<td>5 (S, VP, NP)</td>
<td>6 (S, NP) (S(?!)))</td>
</tr>
</tbody>
</table>

**Mid = 2**
Apparently the sentence is ambiguous for the grammar: (as the grammar overgenerates)
Ambiguity

No subject-verb agreement, and *poison* used as an intransitive verb