Assignment 3 out today: due next Wednesday

Quiz

**Context free grammar**

\[ S \rightarrow NP \ VP \]

- **left hand side**: (single symbol)
- **right hand side**: (one or more symbols)

**Formally...**

- \( G = (NT, T, P, S) \)
- \( NT \): finite set of nonterminal symbols
- \( T \): finite set of terminal symbols, \( NT \) and \( T \) are disjoint
- \( P \): finite set of productions of the form \( A \rightarrow \alpha \), \( A \in NT \) and \( \alpha \in (T \cup NT)^* \)
- \( S \in NT \): start symbol
**CFG: Example**

Many possible CFGs for English, here is an example (fragment):

- $S \rightarrow NP \ VP$
- $VP \rightarrow V \ NP$
- $NP \rightarrow \text{DetP} \ N \mid \text{DetP} \ AdjP \ N$
- $AdjP \rightarrow \text{Adj} \mid \text{Adv} \ AdjP$
- $N \rightarrow \text{boy} \mid \text{girl}$
- $V \rightarrow \text{sees} \mid \text{likes}$
- $\text{Adj} \rightarrow \text{big} \mid \text{small}$
- $\text{Adv} \rightarrow \text{very}$
- $\text{DetP} \rightarrow \text{a} \mid \text{the}$

**Derivations in a CFG**

- $S \rightarrow NP \ VP$
- $VP \rightarrow V \ NP$
- $NP \rightarrow \text{DetP} \ N \mid \text{DetP} \ AdjP \ N$
- $AdjP \rightarrow \text{Adj} \mid \text{Adv} \ AdjP$
- $N \rightarrow \text{boy} \mid \text{girl}$
- $V \rightarrow \text{sees} \mid \text{likes}$
- $\text{Adj} \rightarrow \text{big} \mid \text{small}$
- $\text{Adv} \rightarrow \text{very}$
- $\text{DetP} \rightarrow \text{a} \mid \text{the}$

- $S \rightarrow NP \ VP$
- $VP \rightarrow V \ NP$
- $NP \rightarrow \text{DetP} \ N \mid \text{DetP} \ AdjP \ N$
- $AdjP \rightarrow \text{Adj} \mid \text{Adv} \ AdjP$
- $N \rightarrow \text{boy} \mid \text{girl}$
- $V \rightarrow \text{sees} \mid \text{likes}$
- $\text{Adj} \rightarrow \text{big} \mid \text{small}$
- $\text{Adv} \rightarrow \text{very}$
- $\text{DetP} \rightarrow \text{a} \mid \text{the}$

What can we do?
Derivations in a CFG

\[ S \rightarrow \text{NP} \ \text{VP} \]
\[ \text{VP} \rightarrow \text{V} \ \text{NP} \]
\[ \text{NP} \rightarrow \text{DetP} \ \text{N} \ | \ \text{DetP} \ \text{AdjP} \ \text{N} \]
\[ \text{AdjP} \rightarrow \text{Adj} \ | \ \text{Adv AdjP} \]
\[ \text{N} \rightarrow \text{boy} \ | \ \text{girl} \]
\[ \text{V} \rightarrow \text{sees} \ | \ \text{likes} \]
\[ \text{Adj} \rightarrow \text{big} \ | \ \text{small} \]
\[ \text{Adv} \rightarrow \text{very} \]
\[ \text{DetP} \rightarrow \text{a} \ | \ \text{the} \]
Derivations in a CFG

S → NP VP
VP → V NP
NP → DetP N | DetP AdjP N
AdjP → Adj | Adv AdjP
N → boy | girl
V → sees | likes
Adj → big | small
Adv → very
DetP → a | the

the boy likes NP

Derivations in a CFG

S → NP VP
VP → V NP
NP → DetP N | DetP AdjP N
AdjP → Adj | Adv AdjP
N → boy | girl
V → sees | likes
Adj → big | small
Adv → very
DetP → a | the

the boy likes a girl

Derivations of CFGs

String rewriting system: we derive a string

the boy likes a girl

Derivation history shows the constituent tree:

the boy likes a girl

NP VP

NP V NP

DetP VP
Parsing

Parsing is the field of NLP interested in automatically determining the syntactic structure of a sentence.

Parsing can be thought of as determining what sentences are “valid” English sentences.

As a byproduct, we often can get the structure.

Given a CFG and a sentence, determine the possible parse tree(s):

```
I eat sushi with tuna
```

What parse trees are possible for this sentence?

How did you do it?

What if the grammar is much larger?

What is the difference between these parses?

How can we decide between these?
A Simple PCFG

Probabilities!

<table>
<thead>
<tr>
<th>Rule</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → NP VP</td>
<td>1.0</td>
</tr>
<tr>
<td>NP → NP PP</td>
<td>0.4</td>
</tr>
<tr>
<td>VP → V NP</td>
<td>0.7</td>
</tr>
<tr>
<td>VP → VP PP</td>
<td>0.3</td>
</tr>
<tr>
<td>PP → P NP</td>
<td>1.0</td>
</tr>
<tr>
<td>P → with</td>
<td>1.0</td>
</tr>
<tr>
<td>V → saw</td>
<td>1.0</td>
</tr>
<tr>
<td>NP → astronomers</td>
<td>0.1</td>
</tr>
<tr>
<td>NP → ears</td>
<td>0.18</td>
</tr>
<tr>
<td>NP → saw</td>
<td>0.04</td>
</tr>
<tr>
<td>NP → stars</td>
<td>0.18</td>
</tr>
<tr>
<td>NP → telescope</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Just like n-gram language modeling, PCFGs break the sentence generation process into smaller steps/probabilities.

The probability of a parse is the product of the PCFG rules.

What are the different interpretations here? Which do you think is more likely?
Parsing problems

Pick a model
- e.g. CFG, PCFG, ...

Train (or learn) a model
- What CFG/PCFG rules should I use?
- Parameters (e.g. PCFG probabilities)?
- What kind of data do we have?

Parsing
- Determine the parse tree(s) given a sentence

PCFG: Training

If we have example parsed sentences, how can we learn a set of PCFGs?

Tree Bank

Supervised PCFG Training

English

0.9
0.1
0.6
0.4
0.3
1.0
0.7
0.3
0.2
0.6
0.4
0.3
0.1

Extracting the rules

What CFG rules occur in this tree?

I eat sushi with tuna

<table>
<thead>
<tr>
<th>NP</th>
<th>VP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRP</td>
<td>V N</td>
</tr>
<tr>
<td>S</td>
<td>NP VP</td>
</tr>
<tr>
<td>NP</td>
<td>PRP</td>
</tr>
<tr>
<td>PRP</td>
<td>I</td>
</tr>
<tr>
<td>VP</td>
<td>V NP</td>
</tr>
<tr>
<td>V</td>
<td>eat</td>
</tr>
<tr>
<td>NP</td>
<td>N PP</td>
</tr>
<tr>
<td>N</td>
<td>sushi</td>
</tr>
<tr>
<td>PP</td>
<td>IN N</td>
</tr>
<tr>
<td>IN</td>
<td>with</td>
</tr>
<tr>
<td>N</td>
<td>tuna</td>
</tr>
</tbody>
</table>

Estimating PCFG Probabilities

We can extract the rules from the trees

| S | NP VP |
| NP | PRP |
| PRP | I |
| VP | V NP |
| V | eat |
| NP | N PP |
| N | sushi |
| ... |

How do we go from the extracted CFG rules to PCFG rules?
### Estimating PCFG Probabilities

Extract the rules from the trees

Calculate the probabilities using MLE

\[ p(\alpha \rightarrow \beta | \alpha) = \frac{\text{count}(\alpha \rightarrow \beta)}{\text{count}(\alpha \rightarrow \gamma)} \]

\[ P(\alpha \rightarrow \beta | \alpha) = \frac{\text{count}(\alpha \rightarrow \beta)}{\text{count}(\alpha)} \]

### Estimating PCFG Probabilities

<table>
<thead>
<tr>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>S\rightarrow\text{NP VP}</td>
</tr>
<tr>
<td>S\rightarrow\text{V NP}</td>
</tr>
<tr>
<td>S\rightarrow\text{VP PP}</td>
</tr>
<tr>
<td>\text{NP}\rightarrow\text{N}</td>
</tr>
<tr>
<td>\text{NP}\rightarrow\text{NP PP}</td>
</tr>
<tr>
<td>\text{NP}\rightarrow\text{DT N}</td>
</tr>
</tbody>
</table>

\[ P(\text{S} \rightarrow \text{V NP}) = ? \]

\[ P(\text{S} \rightarrow \text{V NP} | \text{S}) = \frac{\text{count}(\text{S} \rightarrow \text{V NP})}{\text{count}(\text{S})} = \frac{3}{15} \]

### Grammar Equivalence

**Weak equivalence:** grammars generate the same set of strings
- Grammar 1: NP\rightarrow\text{DetP N} and \text{DetP} \rightarrow a \mid \text{the}
- Grammar 2: NP\rightarrow a N \mid \text{the N}

**Strong equivalence:** grammars have the same set of derivation trees
- With CFGs, possible only with useless rules
- Grammar 2: NP\rightarrow a N \mid \text{the N}
- Grammar 3: NP\rightarrow a N \mid \text{the N, DetP} \rightarrow \text{many}
Normal Forms

There are weakly equivalent normal forms (Chomsky Normal Form, Greibach Normal Form)

A CFG is in Chomsky Normal Form (CNF) if all productions are of one of two forms:
- $A \rightarrow BC$ with $A, B, C$ nonterminals
- $A \rightarrow a$, with $A$ a nonterminal and $a$ a terminal

Every CFG has a weakly equivalent CFG in CNF

CNF Grammar

<table>
<thead>
<tr>
<th>Chomsky Normal Form</th>
<th>Probabilistic Grammar Conversion</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow VP$</td>
<td>$S \rightarrow NP \ VP$</td>
</tr>
<tr>
<td>$VP \rightarrow VB \ NP$</td>
<td>$S \rightarrow NP \ VP$</td>
</tr>
<tr>
<td>$NP \rightarrow DT \ NN$</td>
<td>$S \rightarrow NP \ VP$</td>
</tr>
<tr>
<td>$NN \rightarrow VP$</td>
<td>$NP \rightarrow DT \ NN$</td>
</tr>
<tr>
<td>$NP \rightarrow NP \ PP$</td>
<td>$NP \rightarrow NP \ PP$</td>
</tr>
<tr>
<td>$PP \rightarrow IN \ NP$</td>
<td>$PP \rightarrow IN \ NP$</td>
</tr>
<tr>
<td>$IN \rightarrow \ with$</td>
<td>$DT \rightarrow \ the$</td>
</tr>
<tr>
<td>$VB \rightarrow \ film$</td>
<td>$IN \rightarrow \ with$</td>
</tr>
<tr>
<td>$VB \rightarrow \ trust$</td>
<td>$VB \rightarrow \ film$</td>
</tr>
<tr>
<td>$NN \rightarrow \ man$</td>
<td>$VB \rightarrow \ trust$</td>
</tr>
<tr>
<td>$NN \rightarrow \ film$</td>
<td>$NN \rightarrow \ man$</td>
</tr>
<tr>
<td>$NN \rightarrow \ trust$</td>
<td>$NN \rightarrow \ film$</td>
</tr>
</tbody>
</table>

Grammar questions

Can we determine if a sentence is grammatical?

Given a sentence, can we determine the syntactic structure?

Can we determine how likely a sentence is to be grammatical? to be an English sentence?

Can we generate candidate, grammatical sentences?