Welcome to the Statistical Methods of Language Technologyb SoSe21 course

Dr. Seid Muhie Yimam

- Email: yimam@informatik.uni-hamburg.de (mailto:yimam@informatik.uni-hamburg.de)
- Office: Informatikum, F-415

Dr. Özge Alaçam

- Email: alacam@informatik.uni-hamburg.de (mailto:alacam@informatik.uni-hamburg.de)
- Office: Informatikum, F-435

Topic of this week;

In this first practice class, we are going to focus on two main topics, which will be useful to complete the assignment;

- CFGs and PCFG parsing
- Parseval

Deadline: 09/07 June

In class Exercises

Problem 7.1 CFGs and PCFG parsing

a) Convert this CFG to CNF. Capital letters are non-terminals and lowercase letters are terminals, S is the start symbol.

$$
\begin{aligned}
S &\rightarrow A\ B\ C \mid S\ D \\
A &\rightarrow D\ D \mid \epsilon \\
B &\rightarrow F \mid C\ C\ C \\
C &\rightarrow a\ c \mid a\ d \mid a\ e \mid \epsilon \\
D &\rightarrow C\ d\ d \mid d\ e \\
F &\rightarrow \epsilon
\end{aligned}
$$

Context Free grammar to Chomsky normal form (CNF) conversion

CONVERSION TO CHOMSKY NORMAL FORM

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

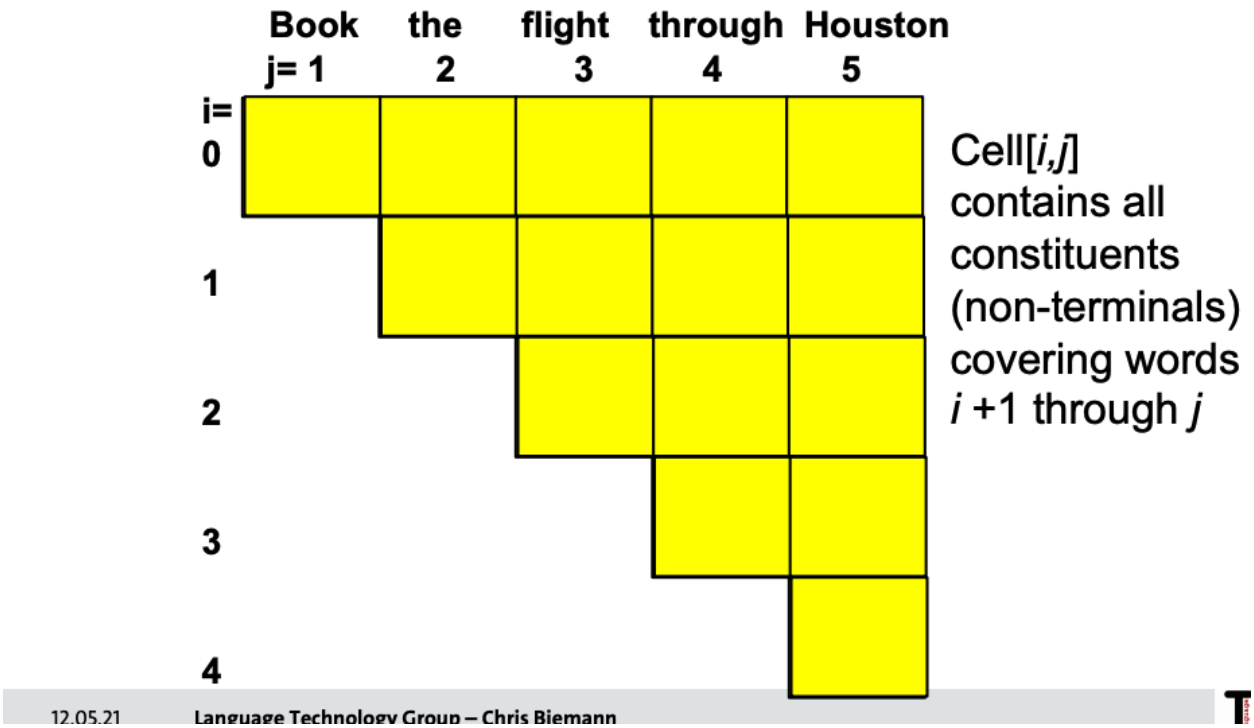
1. Introduce a new start symbol S_0, add rule $S_0 \rightarrow S$ (S=old start symbol)
2. Eliminate all ε rules of the form $A \rightarrow \varepsilon$ ($A \neq S_0$): remove rule and split rules containing A on the RHS in all versions, with and without A's. For rules $B \rightarrow A$, replace A with ε if B has not been through this step yet, otherwise eliminate $B \rightarrow A$.
3. Eliminate all unit rules $A \rightarrow B$, by adding all $B \rightarrow R_i$ to $A \rightarrow R_i$ where R_i is not a unit rule. If R_i is a unit rule add all $R_i \rightarrow K_i$ to A ($A \rightarrow K_i$) where K_i is not a unit rule. Continue this process for all following unit-rules, until we observe a unit rule we have seen in the cleaning step. Then eliminate $A \rightarrow B$.
4. Clean up remaining rules: For $A \rightarrow R_1, R_2, .. R_n$ (n>2, R_i terminals or non-terminals), create a chain $\{A \rightarrow R_1 A_1, A_1 \rightarrow R_2 A_2 \ldots A_{n-2} \rightarrow R_{n-1} R_n\}$. For all R_i that are terminals, create a lexicon rule and replace R_i with its LHS.
5. If $S_0 \rightarrow C$ remains, set C as start symbol.

b) Parse the sentence "I see the man with the telescope using the bottom up **CYK" algorithm (convert to CNF)

Assume the following grammar:

$$
\begin{aligned}
S &\rightarrow NP\ VP \mid NP \\
VP &\rightarrow V\ NP \mid V\ NP\ PP \\
NP &\rightarrow NP\ PP \mid D\ N \mid R \\
PP &\rightarrow P\ NP \\
P &\rightarrow with \mid in \mid of \\
V &\rightarrow see \mid take \mid make \\
N &\rightarrow man \mid telescope \\
D &\rightarrow the \mid an \mid a \\
R &\rightarrow I
\end{aligned}
$$

CYK PARSER



c) Given the following grammar and lexicon rules, parse the sentence ``fruit flies like a banana'' with probabilistic CYK.

- Grammar

$$
\begin{aligned}
S &\rightarrow NP\ VP & 1.0\\
NP &\rightarrow D\ N & 0.5\\
NP &\rightarrow N\ N & 0.3\\
NP &\rightarrow NP\ PP & 0.2\\
VP &\rightarrow V\ NP & 0.6\\
VP &\rightarrow V\ PP & 0.4\\
PP &\rightarrow P\ NP & 1.0
\end{aligned}
$$

- Lexicon

$$
\begin{aligned}
N &\rightarrow fruit & 0.2\\
V &\rightarrow flies & 0.1\\
N &\rightarrow flies & 0.1\\
V &\rightarrow like & 0.3\\
P &\rightarrow like & 0.2\\
D &\rightarrow a & 0.4\\
D &\rightarrow the & 0.5\\
N &\rightarrow banana & 0.1\\
N &\rightarrow time & 0.1
\end{aligned}
$$

Hints from lecture Slides:

PROBABILISTIC CONVERSION TO CNF

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

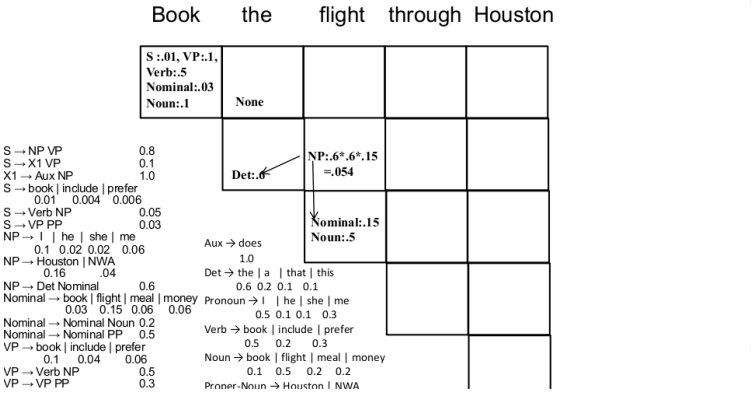
Original Grammar		Chomsky Normal Form	
S → NP VP	0.8	S → NP VP	0.8
S → Aux NP VP	0.1	S → X1 VP	0.1
		X1 → Aux NP	1.0
S → VP	0.1	S → book | include | prefer 0.01 0.004 0.006	
		S → Verb NP	0.05
		S → VP PP	0.03
NP → Pronoun	0.2	NP → I | he | she | me 0.1 0.02 0.02 0.06	
NP → Proper-Noun	0.2	NP → Houston | NWA 0.16 .04	
NP → Det Nominal	0.6	NP → Det Nominal	0.6
Nominal → Noun	0.3	Nominal → book | flight | meal | money 0.03 0.15 0.06 0.06	
Nominal → Nominal Noun	0.2	Nominal → Nominal Noun	0.2
Nominal → Nominal PP	0.5	Nominal → Nominal PP	0.5
VP → Verb	0.2	VP → book | include | prefer 0.1 0.04 0.06	
VP → Verb NP	0.5	VP → Verb NP	0.5
VP → VP PP	0.3	VP → VP PP	0.3
PP → Prep NP	1.0	PP → Prep NP	1.0

Hints from lecture Slides:

First Steps

PROBABILISTIC CYK PARSING

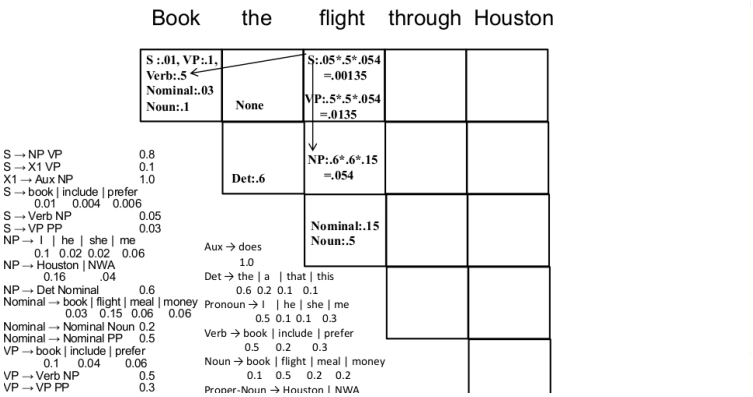
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Next Step

PROBABILISTIC CYK PARSING

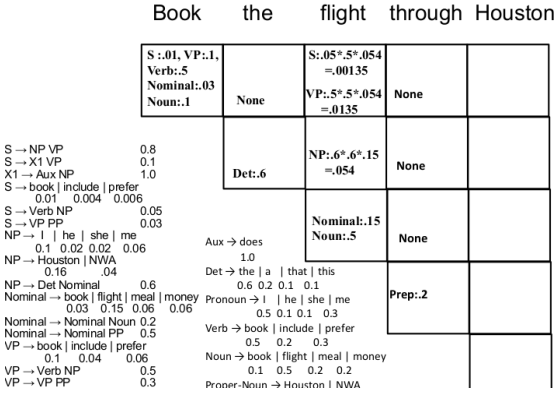
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Next Step

PROBABILISTIC CYK PARSING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Final Step

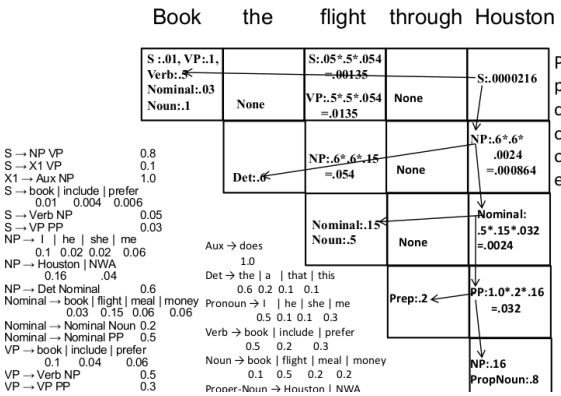
SIMPLE PCFG FOR A SUBSET OF ENGLISH

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Grammar	Prob.		Lexicon
S → NP VP	0.8		Det → the | a | that | this
S → Aux NP VP	0.1	+ 1.0	0.6 0.2 0.1 0.1
S → VP	0.1		Noun → book | flight | meal | money
NP → Pronoun	0.2		0.1 0.5 0.2 0.2
NP → Proper-Noun	0.2	+ 1.0	Verb → book | include | prefer
NP → Det Nominal	0.6		0.5 0.2 0.3
Nominal → Noun	0.3		Pronoun → I | he | she | me
Nominal → Nominal Noun	0.2	+ 1.0	0.5 0.1 0.1 0.3
Nominal → Nominal PP	0.5		Proper-Noun → Houston | NWA
VP → Verb	0.2		0.8 0.2
VP → Verb NP	0.5	+ 1.0	Aux → does
VP → VP PP	0.3		1.0
PP → Prep NP	1.0		Prep → from | to | on | near | through
			0.25 0.25 0.1 0.2 0.2

PROBABILISTIC CYK PARSING

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

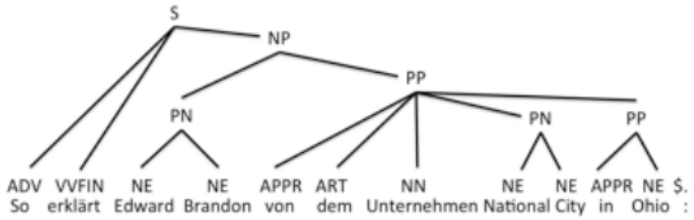


Type *Markdown* and LaTeX: α^2

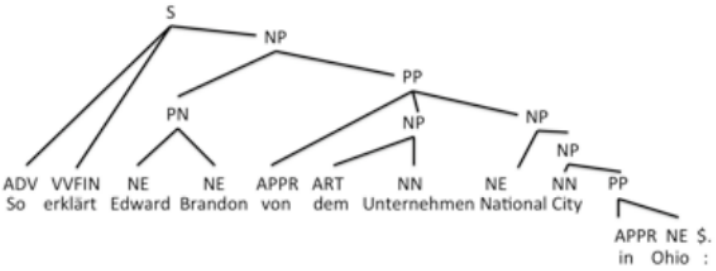
Problem 7.2 Parse Evaluation

a) Compute Precision, Recall and F1 value for the following two parses, labeled and unlabeled.

System:



Gold:



b) Convert the trees in a) into the Penn Treebank bracketed format.

Hints from Lecture Slides:

FORMAT

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Every production rule is represented by

- (
- left hand side
- sequence of right hand side symbols
  - non-terminals expanded by production rule
  - terminals
-)

Traces: -NONE- and trace-number

```
( (S
  (NP-SBJ (DT The) (NNP Illinois) (NNP Supreme) (NNP Court) )
  (VP (VBD ordered)
   (NP-1 (DT the) (NN commission) )
   (S
    (NP-SBJ (-NONE- *-1) )
    (VP (TO to)
     (VP
      (VP (VB audit)
       (NP
        (NP (NNP Commonwealth) (NNP Edison) (POS 's) )
        (NN construction) (NNS expenses) ))
      (CC and)
      (VP (VB refund)
       (NP (DT any) (JJ unreasonable) (NNS expenses) ))))))
(. .) ))
```

c) Familiarize with the Stanford CoreNLP Parser https://corenlp.run/ (https://corenlp.run/) or http://nlp.stanford.edu:8080/parser/ (http://nlp.stanford.edu:8080/parser/) . Parse a sentence and draw the tree (using constituency parse). Does your sentence make sense?

In [ ]: