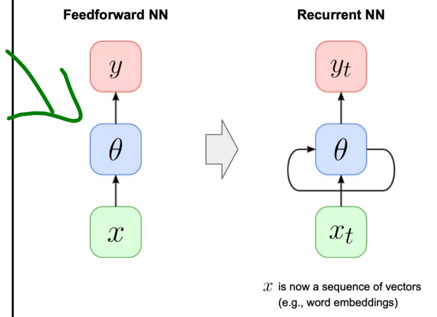
CS4248: Natural Language Processing

Lecture 9 — Trees

Recap of Week 08

Recurrent Neural Network — Basic Idea



Core concept of RNNs: **Hidden State**

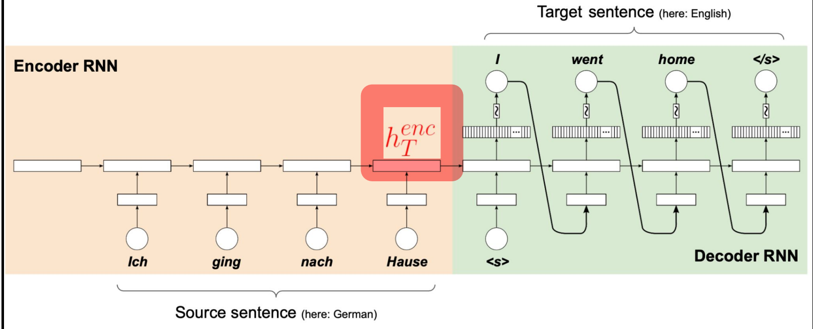
- Additional vector incorporated into the network
- Commonly holds the last output of the hidden layer
  ➜ size of hidden state = size of hidden layer
- Randomly initialized, and to be tuned through training (➜ backpropagation)
- Basic recurrent formula:

$$h_t = f_\theta(h_{t-1}, x_t)$$

hidden state of time step $t-1$ — input vector at time step t

20

Encoder-Decoder (Sutskever et al.; 2014)



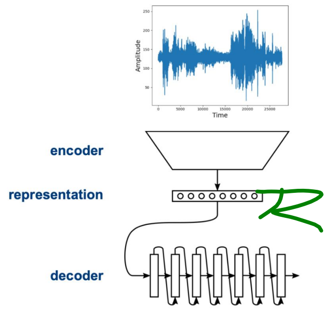
55

Encoder-Decoder Architecture

- Basic 2-component setup

  (1) Encoder
  - Learns function that maps context into a fixed-size vector representation c
  - Encoder architecture depending on context
    (e.g., CNN for images, RNN for text)

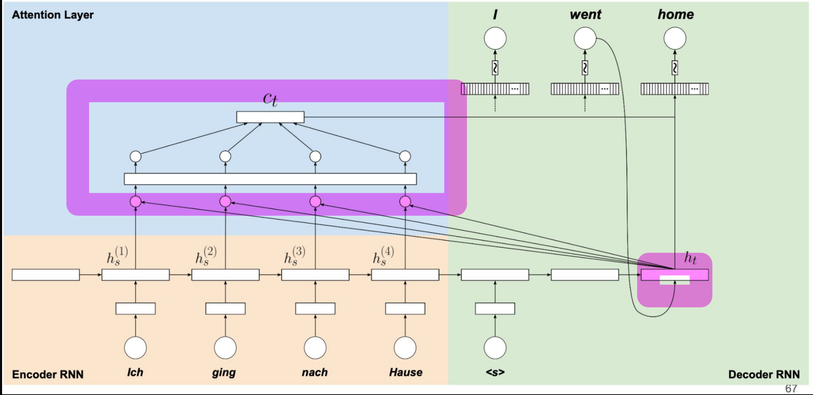
  (2) Decoder
  - Language model using c to output sequence of words
  - In the following: RNN-based Decoder



"Back off man, I'm a scientist."

50

Attention — Walkthrough



67

Announcements

Assignments

- Assignment 1: Graders are examining your requests.
- Assignment 1: We are also processing our requests on possible academic dishonesty.
- Assignment 3 out: Theory and Practice, 3 problems.

Projects

- Intermediate Updates to be disseminated soon. 5% of your total grade, uniformly distributed.
- First TEAMMATES intra-peer evaluation has been completed. Please check your own results.
- Min and Chris will be announcing a sign-up sheet for teams to optionally consult us on their project.
- Teams can be nominated or self-nominate for STePS (Week 13 Wed) for poster styled presentation instead.

Outline

- **Syntactic Parsing**
  - **Quick recap: RegEx**
  - Context-free grammars
  - Structural Ambiguity
  - Chomsky Normal Form
- CYK Parsing Algorithm
  - Base membership algorithm
  - Find all parse trees with backtracking
  - Probabilistic parsing
  - Evaluation of parser

Throwback — Regular Expression

- Equivalence
  - Regular Expressions describe **Regular Languages**
    (most restricted types of languages w.r.t Chomsky Hierarchy)
  - Regular Language = language accepted by a FSA

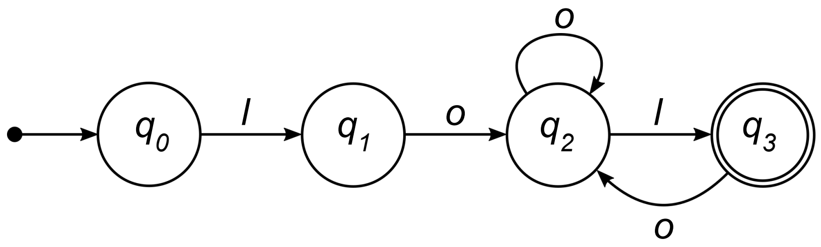
Regular Expression

l(o+l)+

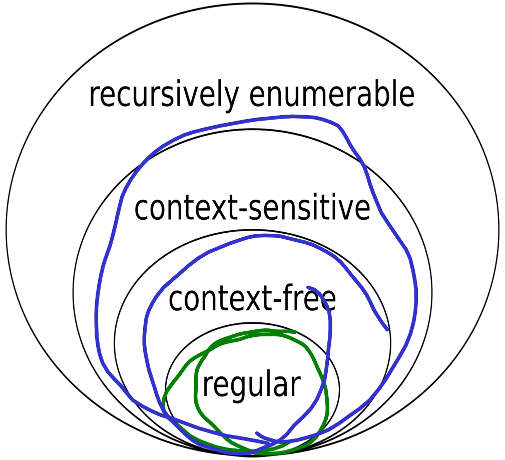
Regular Language

{lol, loool, lolol, looolol, …}

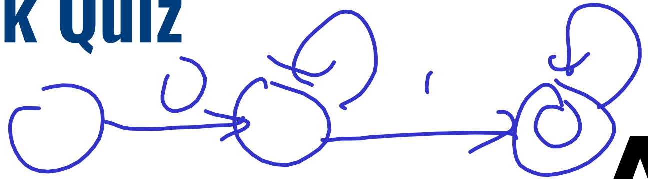
Example: FSA that accepts the Regular Language described by the Regular Expression **l(o+l)+**



Chomsky Hierarchy
(Source: Wikipedia)

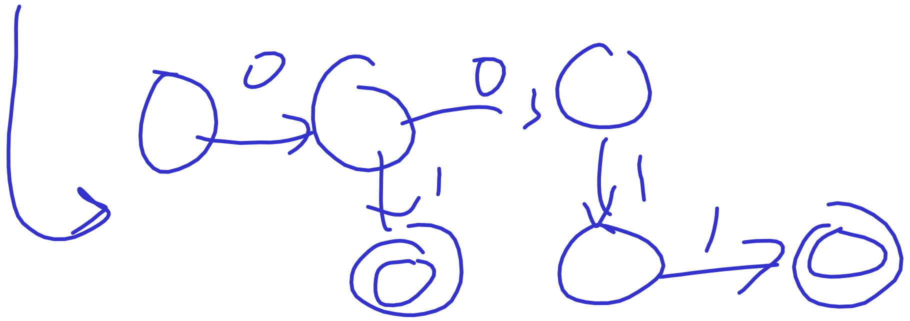


🏃‍♀️🏃🏃‍♂️ Quick Quiz



Can you build an FSA / RegEx that **accepts** the language below?

$$\{\epsilon, 01, 0011, 000111\}$$



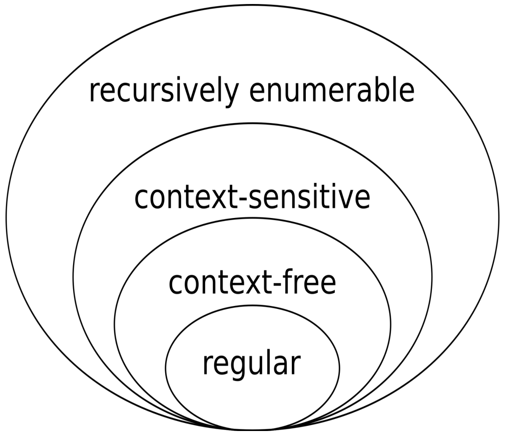
A Yes

[✓ B No]

C Undecidable

D RegEx? What's a RegEx?

Regular Expressions — Limitations



- Not all languages can be described using RegEx
  - Example:

$$\{\underline{0^n}\,\underline{1^n} \mid n \geq 0\} = \{\epsilon, 01, 0011, 000111, 00001111, \ldots\}$$

➜ Natural Language is not a Regular Language

- Natural language allows for nested structures (**center embeddings**)

The food was delicious

The food Alice cooked was delicious

The food Alice the sister of Bob cooked was delicious

Syntax & Constituency

- Important questions
  - How are words combined to form phrases?
  - How are phrases combined to form longer phrases?
  - How are phrases are combined to form sentences?

How meaning is mapped onto what language structures?

- Important concept: constituency = phrase structure
  - Constituent = group of words that behave as a single unit

Constituents

- Constituent — Definition
  - Group of words that behaves as a single unit or phrase
    (by default: individual words are constituents, but there are exceptions)
  - Sentences can be described as a hierarchical structure of constituents (in a bit: parse trees)

- Question: How do we know a group of words forms a constituent?
  - Handwavy answer: Group of words that "make sense" on its own

"She heard a loud shot from the bank during the time of the robbery."

a loud shot

a loud shot from the

a loud shot from the bank

  - Formal answer: **Constituency Tests**

Constituency Tests (some examples)

1 • Topicalization
  - Only a constituent can be moved to different locations in a sentence

    “They met ***at 8 pm*** *for dinner.” — “They met for dinner* ***at 8 pm****.” — “****At 8 pm****, they met for dinner.”*

2 • Proform substitution
  - Only a constituent can be substituted with a proform like *it*, *that*, *them*, *then*, *there*, etc.

    “Chris went ***back to Germany****.”* ➜ *“Chris went there.”*

3 • Fragment Answers
  - Only a constituent can answer a question while retaining the meaning of the original sentence.

    “Alice was hit by ***the green car****.” — Q: “What hit Alice?”* ➜ *“The green car.”*

🏃‍♀️🏃🏃 Quick Quiz

Which is **not a constituent** in the sentence below?

"All students learned about syntactic parsing this week"

A *All students*

B *learned about syntactic parsing*

C *this week*

D *learned about* *

Outline

- **Syntactic Parsing**
  - Quick recap: RegEx
  - **Context-free grammars**
  - Structural Ambiguity
  - Chomsky Normal Form
- CYK Parsing Algorithm
  - Base membership algorithm
  - Find all parse trees with backtracking
  - Probabilistic parsing
  - Evaluation of parser

Context-Free Grammars (CFGs)

- Context-Free Grammars
  - Most common way to capture constituency and ordering ➜ good fit for natural language!
    (in fact, context-free grammars were first used to study human languages to describe the structure of sentences)
  - Define what meaningful constituents are and how a constituent is formed out of other constituents
  - More powerful than RegExs as they can express recursive structure
    (in contrast, context free grammars can describe regular languages)

- Example

special start symbol $\longrightarrow$

$$\underbrace{\begin{array}{rcl} S & \xrightarrow{\text{rewrite}} & \text{NP VP} \\ \text{NP} & \rightarrow & \text{Det Noun} \\ \text{VP} & \rightarrow & \text{Verb NP} \\ \text{Det} & \rightarrow & \mathit{a} \mid \mathit{the} \\ \text{Noun} & \rightarrow & \mathit{man} \mid \mathit{meal} \mid \mathit{flight} \\ \text{Verb} & \rightarrow & \mathit{saw} \mid \mathit{booked} \end{array}}_{\text{set of } \textbf{rules} \text{ or } \textbf{productions}}$$

✓ **Non-terminal symbols**
- Symbols that can be replaced according to rules
- For natural language grammars: phrase names, part of speech

✓ **Terminal symbols**
- May be the output of a rule; cannot be changed/replaced further
- For natural language grammars: words/tokens

Context-Free Grammars (CFGs)

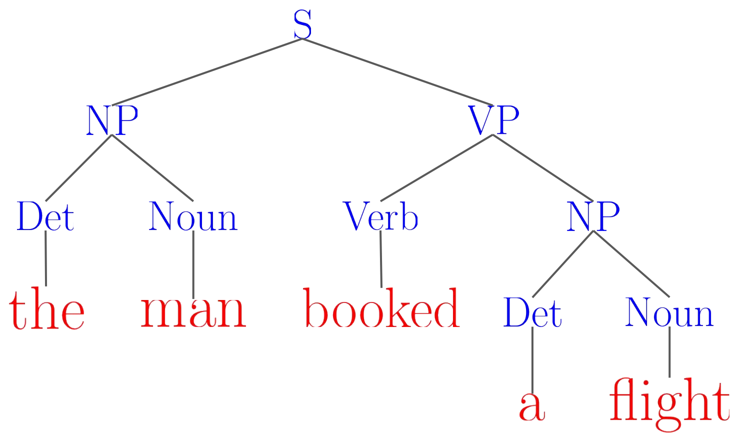
- Application of rules — example

$$\begin{aligned} \text{S} &\to \text{NP VP} \\ \text{NP} &\to \text{Det Noun} \\ \text{VP} &\to \text{Verb NP} \\ \text{Det} &\to \textit{a} \mid \textit{the} \\ \text{Noun} &\to \textit{man} \mid \textit{meal} \mid \textit{flight} \\ \text{Verb} &\to \textit{saw} \mid \textit{booked} \end{aligned}$$

$$\begin{aligned} \text{S} &\to \text{NP VP} \\ &\to \text{Det Noun VP} \\ &\to \textit{the}\ \text{Noun VP} \\ &\to \textit{the man}\ \text{VP} \\ &\to \textit{the man}\ \text{Verb NP} \\ &\to \textit{the man booked}\ \text{NP} \\ &\to \textit{the man booked}\ \text{Det Noun} \\ &\to \textit{the man booked a}\ \text{Noun} \\ &\to \textit{the man booked a flight} \end{aligned}$$

Derivation: sequence of rules/productions used to generate a string of words

Visualization as **Parse Tree**



CFG — Formal Definition

- A CFG is a 4-tuple $\langle N, \Sigma, R, S \rangle$
  - N — set of non-terminal symbols
  - Σ — set of terminal symbols
  - R — set of rules

Allowed format for all rules

$A \rightarrow \alpha$ with $A \in N,\ \alpha \in (N \cup \Sigma)$

  - S — start symbol

Example

$N = \{\mathrm{NP},\ \mathrm{VP},\ \mathrm{Det},\ \mathrm{Noun},\ \mathrm{Verb}\}$

$\Sigma = \{a,\ the,\ man,\ meal,\ flight,\ saw,\ booked\}$

$$\begin{aligned} \mathrm{S} &\rightarrow \mathrm{NP\ VP} \\ \mathrm{NP} &\rightarrow \mathrm{Det\ Noun} \\ \mathrm{VP} &\rightarrow \mathrm{Verb\ NP} \\ \mathrm{Det} &\rightarrow a \mid the \\ &\ldots \end{aligned}$$

🏃‍♀️🏃🏃‍♂️ Quick Quiz

✏️ What part of this specification makes this context-free?

- A CFG is a 4-tuple
  - set of non-terminal symbols *N*
  - set of terminal symbols $\sum$
  - set of rules $R : A \to \beta$
  - start symbol *S*

A N

B Σ

C [✓ $R\text{: } A \to \beta$]

D $R\text{: } \beta \in (\Sigma, N)$

CFG — Running Example

$\boxed{}$ A B C -> A D C
X B C -> X F C

$$\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{S} &\rightarrow \text{Aux NP VP} \\
\text{S} &\rightarrow \text{VP} \\
\text{NP} &\rightarrow \text{Pronoun} \\
\text{NP} &\rightarrow \text{ProperNoun} \\
\text{NP} &\rightarrow \text{Det Nominal} \\
\text{Nominal} &\rightarrow \text{Noun} \\
\text{Nominal} &\rightarrow \text{Nominal Noun} \\
\text{Nominal} &\rightarrow \text{Nominal PP} \\
\text{VP} &\rightarrow \text{Verb} \\
\text{VP} &\rightarrow \text{Verb NP} \\
\text{VP} &\rightarrow \text{Verb NP PP} \\
\text{VP} &\rightarrow \text{Verb PP} \\
\text{VP} &\rightarrow \text{VP PP} \\
\text{PP} &\rightarrow \text{Prep NP}
\end{aligned}$$

$$\begin{aligned}
\text{Det} &\rightarrow the \mid a \mid that \mid this \\
\text{Noun} &\rightarrow book \mid flight \mid meal \mid money \\
\text{Verb} &\rightarrow book \mid include \mid prefer \\
\text{Pronoun} &\rightarrow I \mid she \mid he \mid me \\
\text{ProperNoun} &\rightarrow Singapore \mid Frankfurt \mid SIA \\
\text{Aux} &\rightarrow do \mid does \mid did \\
\text{Prep} &\rightarrow from \mid to \mid on \mid near \mid through
\end{aligned}$$

Important requirements to make it a CFG

- Only single non terminals on the ~~right~~ left-hand side

➜ Application of a rule does not depend on context

🏃‍♀️🏃🏃 Quick Quiz

How many **issues** make the grammar below **not** a CFG?

S → NP VP
S → S CC S
NP → Pronoun
NP → ProperNoun
NP → Det Nominal
Nominal → *the* Noun
Nominal → Nominal Noun
VP NP → Verb NP
VP → Verb
VP → Verb NP PP
VP → Verb PP
the → Det
...

A [✓ 1]

B [✓ 2]

C 3

D 4

Outline

- **Syntactic Parsing**
  - Quick recap: RegEx
  - Context-free grammars
  - **Structural Ambiguity**
  - Chomsky Normal Form
- CYK Parsing Algorithm
  - Base membership algorithm
  - Find all parse trees with backtracking
  - Probabilistic parsing
  - Evaluation of parser

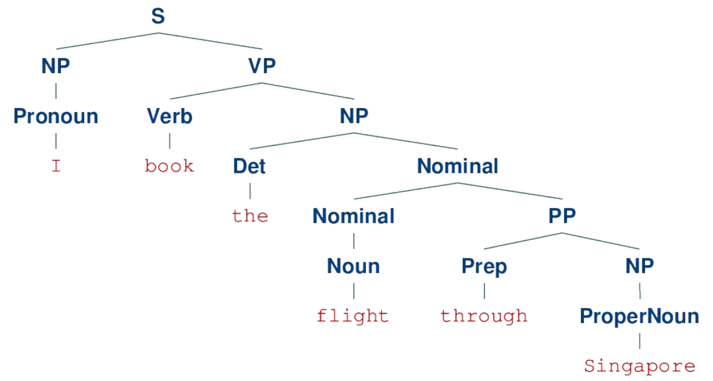
Ambiguity

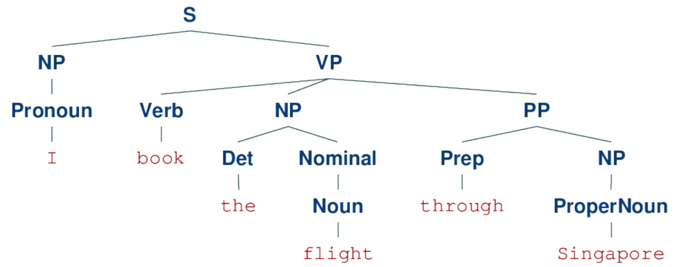
- Ambiguity of Natural Language
  - Common: multiple ways to interpret a sentence
  - Different interpretation ➜ different meaning

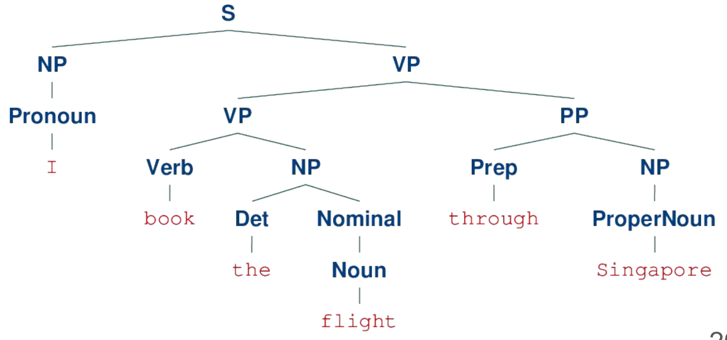
➜ Structural Ambiguity

- A grammar can assign more than one parse to a sentence
- Example (using our toy grammar):

  "I book the flight through Singapore"







🤷‍♂️🧑‍🤷🤷‍♀️ Structural Ambiguity

- Two common types of Structural Ambiguity

(1) **Attachment Ambiguity**

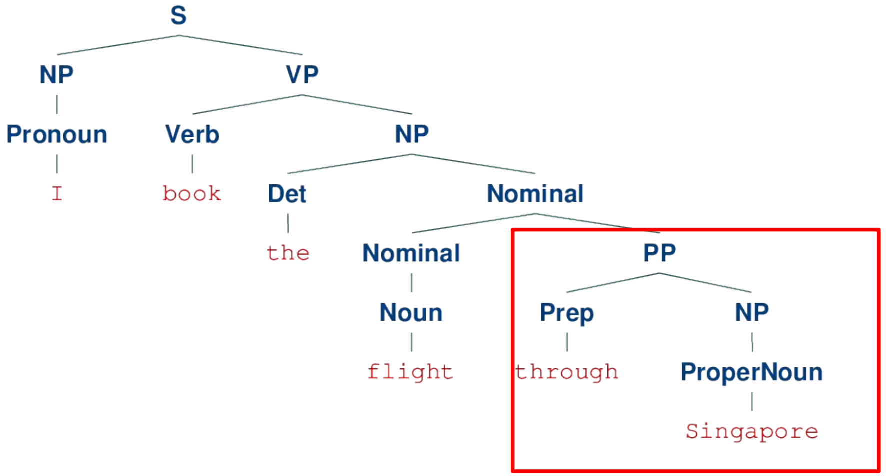
- A particular constituent can be attached to the parse tree at more than one place

(2) **Coordination Ambiguity**

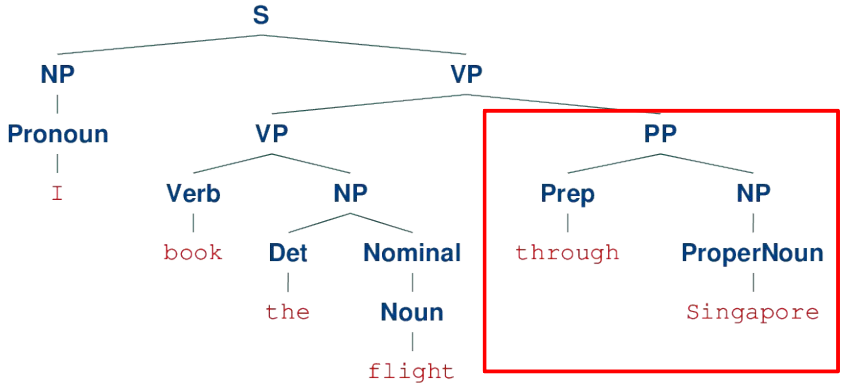
- Phrases can be conjoined by conjunction like *“and”*, *“or”*, *“but”*, *“because”*, *“if”*, etc.
- Different types of conjunctions
(coordinating conjunctions, correlative conjunctions, correlative conjunctions)

🏃‍♀️ 🏃 🏃‍♂️ Which parse is correct?

"I book the flight through Singapore"



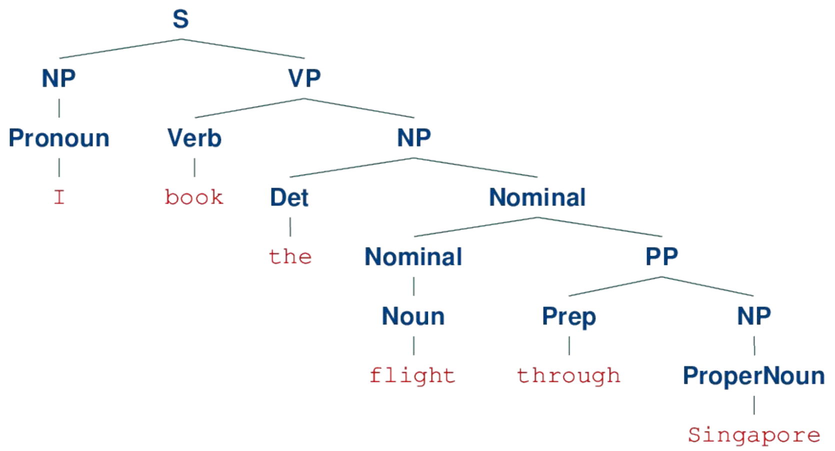
A



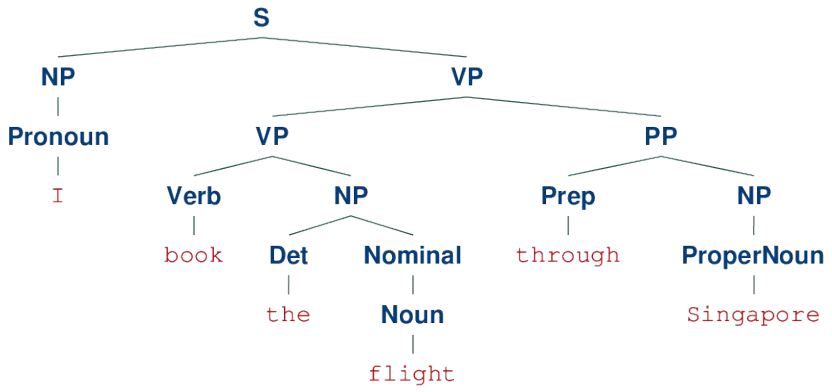
B

Attachment Ambiguity

“I book the flight through Singapore”



“through Singapore” attached to noun phrase related to *“flight”*

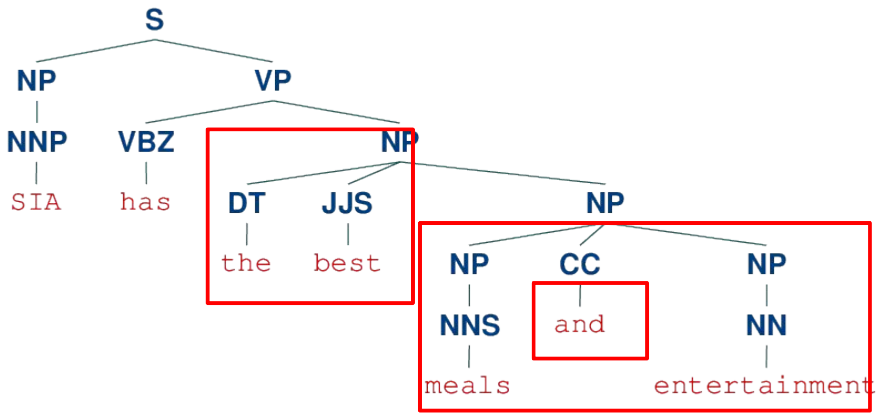


“through Singapore” attached to verb phrase related to *“book”*

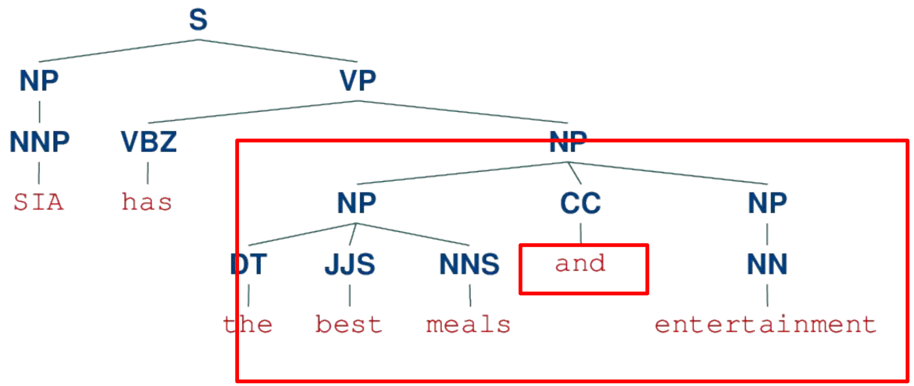
(like booking a flight through an agent)

🤷‍♀️🧑🤷‍♂️ Coordination Ambiguity

"SIA has the best meals and entertainment"



"best" refers to both the meals and the entertainment



"best" refers to only the meals but not the entertainment

Note: This example used a different grammar since our toy grammar does not support conjunctions.

Pre-Lecture Activity for Last Week

- Assigned Task
  - Watch the 9-minute YouTube video linked below
  - Take an ambiguous news headline and explain one strategy mentioned in the video
  - Post a 1-2 sentence answer to the following questions in your Tutorial Group's discussion

The Ling Space:
"How Do We Interpret Sentences? Parsing Strategies"

Side notes:
- This task is meant as a warm-up to provide some context for the next lecture
- No worries if you get lost; we will talk about this in the next lecture
- You can just copy-&-paste others' answers but this won't help you learn better

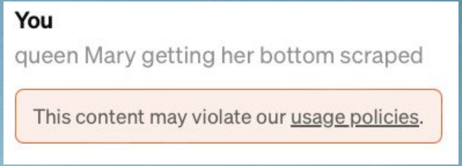
Sisters reunited after 18 years at checkout counter

From my point of view, Late Closure is kind of opposite to the usually used proximity agreement in grammar. Other than sticking to the most recently appeared word before it, this word should stick to another word which appears quite early in the sentence.

E.g. "2 SISTERS REUNITED AFTER 18 YEARS AT CHECKOUT COUNTER". Normally our intuitive parsing will stick "at check COUNTER" with "18 years" which explains the headline as "2 sisters reunited after they spent 18 years at checkout counter", and that is totally wrong. By Late Closure strategy, we can correctly interpret the headline as "2 sisters reunited at the checkout counter after 18 years of not seeing each other" because we stick "at checkout counter" to "reunited" in this case.

Graphics courtesy Dall.E (run 22 Mar 2024)

Queen Mary getting her bottom scraped



You

queen Mary getting her bottom scraped

This content may violate our usage policies.

The principle of minimal attachment suggests that when faced with multiple ways to interpret a sentence, the human language processing system tends to choose the interpretation that involves the fewest assumptions.

For example, the phrase "Queen Mary having bottom scraped" . First we may think queen Mary is a person. In this case, "having bottom scraped" may not make sense. But if then we know "Queen Mary" refers to a ship. "Having bottom scraped"then be a maintenance process.

The principle of minimal attachment might initially lead one to choose the interpretation that fits with the simplest and most direct syntactic structure based on their knowledge and context.

Graphics courtesy Dall.E (Run 22 Mar 2024)

Summary So Far...

- Parsing as a 2-part task

    (1) **Syntactic Parsing**
    - Extract all possible parses for a sentence
    - Typically requires a grammar transformation step
      (“binarization” of grammar to ensure efficient parsing)

    (2) **Syntactic Disambiguation**
    - Score all parses and return the best parse
    - Scores commonly expressed as probability

Outline

- **Syntactic Parsing**
  - Quick recap: RegEx
  - Context-free grammars
  - Structural Ambiguity
  - **Chomsky Normal Form**
- CYK Parsing Algorithm
  - Base membership algorithm
  - Find all parse trees with backtracking
  - Probabilistic parsing
  - Evaluation of parser

Grammar Transformation (for CFGs)

- Important requirement: binarized rules
  - No more than 2 non-terminals on the right-hand side of rules
  - Crucial for a cubic time parsing of CFGs

➜ Common transformation: **Chomsky Normal Form**

- Restrictions on rules compared to general CFG

Allowed format for all rules

$A \rightarrow \alpha$ with $A \in N, \quad \alpha \in (N \cup \Sigma)$

α **is either <u>1 terminal</u> OR <u>2 non-terminals</u>**

🏃‍♀️🏃🏃‍♂️ A Great Way to Fly

Allowed format for all rules

$A \rightarrow \alpha$ with $A \in N$, $\alpha \in (N \cup \Sigma)$

α **is either 1 terminal OR 2 non-terminals**

grammar - all non-terminals (both LHS and RHS)

$S \rightarrow NP\ VP$
$S \rightarrow Aux\ NP\ VP$
$S \rightarrow VP$
$NP \rightarrow Pronoun$
$NP \rightarrow ProperNoun$
$NP \rightarrow Det\ Nominal$
$Nominal \rightarrow Noun$
$Nominal \rightarrow Nominal\ Noun$
$Nominal \rightarrow Nominal\ PP$
$VP \rightarrow Verb$
$VP \rightarrow Verb\ NP$
$VP \rightarrow Verb\ NP\ PP$
$VP \rightarrow Verb\ PP$
$VP \rightarrow VP\ PP$
$PP \rightarrow Prep\ NP$

$Det \rightarrow the \mid a \mid that \mid this$
$Noun \rightarrow book \mid flight \mid meal \mid money$
$Verb \rightarrow book \mid include \mid prefer$
$Pronoun \rightarrow I \mid she \mid he \mid me$
$ProperNoun \rightarrow Singapore \mid Frankfurt \mid SIA$
$Aux \rightarrow do \mid does \mid did$
$Prep \rightarrow from \mid to \mid on \mid near \mid through$

lexicon

<--------- Det -> the
Det -> a
Det -> that
Det -> this (4x)

non-terminal rewritten to a single terminal symbol

Quick Quiz:

- Which rules do not conform to the Chomsky Normal Form?
- How can we transform the grammar to fix this?

Chomsky Normal Form (CNF)

- Two basic transformation steps

(1) Recursive removal of unary rules (and empty rules)

$$\begin{aligned} \text{Nominal} &\rightarrow \text{Noun} \\ \text{Noun} &\rightarrow \textit{book} \mid \textit{flight} \mid \textit{meal} \mid \textit{money} \end{aligned} \quad\Rightarrow\quad \begin{aligned} \text{Nominal} &\rightarrow \textit{book} \mid \textit{flight} \mid \textit{meal} \mid \textit{money} \\ \text{Noun} &\rightarrow \textit{book} \mid \textit{flight} \mid \textit{meal} \mid \textit{money} \end{aligned}$$

(2) Dividing n-ary rules by introducing new non-terminals
(n-ary rule = rule with n > 2 non-terminal on the right-hand side)

$$\text{S} \rightarrow \text{Aux NP VP} \quad\Rightarrow\quad \begin{aligned} S &\rightarrow \boxed{X}\, VP \\ \boxed{X} &\rightarrow Aux\ NP \end{aligned}$$

Toy Grammar in Chomsky Normal Form CNF

$\text{S} \rightarrow \text{NP VP}$
$\text{S} \rightarrow \text{X1 VP}$
$\text{X1} \rightarrow \text{Aux NP}$
$\text{S} \rightarrow \text{Verb NP}$
$\text{S} \rightarrow \text{X2 PP}$
$\text{S} \rightarrow \text{Verb PP}$
$\text{S} \rightarrow \text{VP PP}$
$\text{NP} \rightarrow \text{Det Nominal}$
$\text{Nominal} \rightarrow \text{Nominal Noun}$
$\text{Nominal} \rightarrow \text{Nominal PP}$
$\text{VP} \rightarrow \text{Verb NP}$
$\text{VP} \rightarrow \text{X2 PP}$
$\text{X2} \rightarrow \text{Verb NP}$
$\text{VP} \rightarrow \text{Verb PP}$
$\text{VP} \rightarrow \text{VP PP}$
$\text{PP} \rightarrow \text{Prep NP}$

$\text{Det} \rightarrow the \mid a \mid that \mid this$
$\text{Noun} \rightarrow book \mid flight \mid meal \mid money$
$\text{Verb} \rightarrow book \mid include \mid prefer$
$\text{Pronoun} \rightarrow I \mid she \mid he \mid me$
$\text{PropNoun} \rightarrow Singapore \mid Frankfurt \mid SIA$
$\text{Aux} \rightarrow do \mid does \mid did$
$\text{Prep} \rightarrow from \mid to \mid on \mid near \mid through$
$\text{S} \rightarrow book \mid include \mid prefer$
$\text{VP} \rightarrow book \mid include \mid prefer$
$\text{NP} \rightarrow I \mid she \mid he \mid me$
$\text{NP} \rightarrow Singapore \mid Frankfurt \mid SIA$
$\text{Nominal} \rightarrow book \mid flight \mid meal \mid money$

Allowed format for all rules

$A \rightarrow \alpha$ with $A \in N,\ \ \alpha \in (N \cup \Sigma)$

α **is either 1 terminal OR 2 non-terminals**

CFG to CNF — Summary

- Transformation of a CFG to a CNF
  - Every CFG can be transformed into a **weakly-equivalent** CNF

➜ **Weak equivalence**

- Two grammars generate the same set of sentences (identical expressiveness)
- The derivations generating the same sentences may differ
  (recall that the CNF may introduce additional non-terminals)

(Strong equivalence: identical expressiveness + identical derivations)

Noam Chomsky

Sometimes referred to the father of modern linguistics. Originated many key ideas, including universal grammar and a normal form for context free grammars.

He is a polymath, a very influential scholar for his critique of U.S. foreign policies.

His doctoral thesis impressed George A. Miller, noted psychologist and the originator of WordNet.

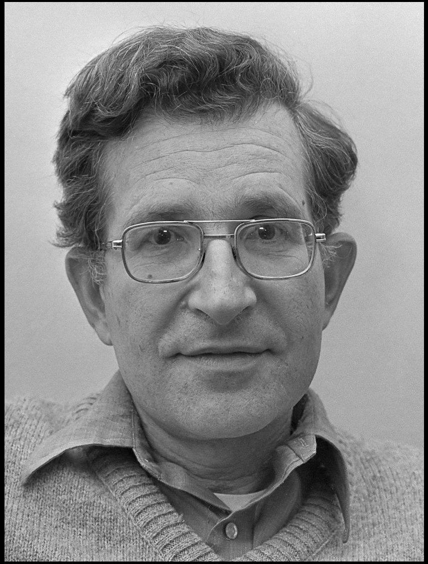
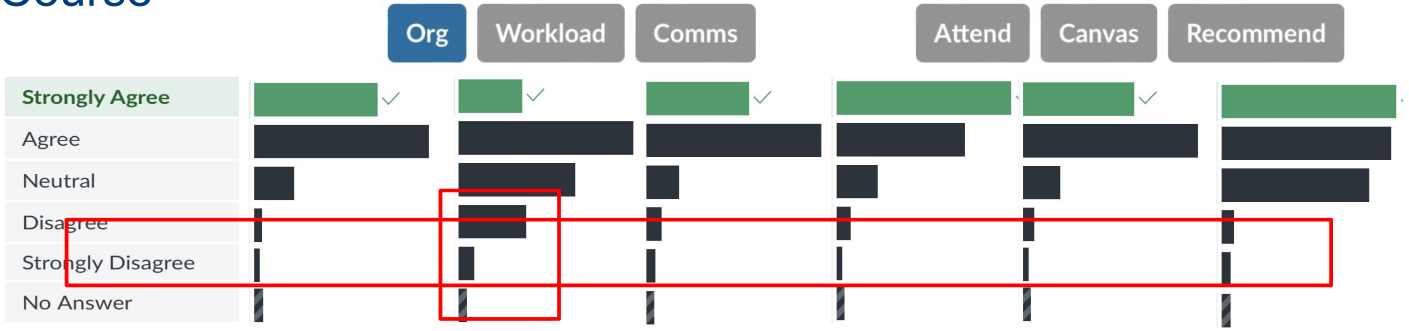


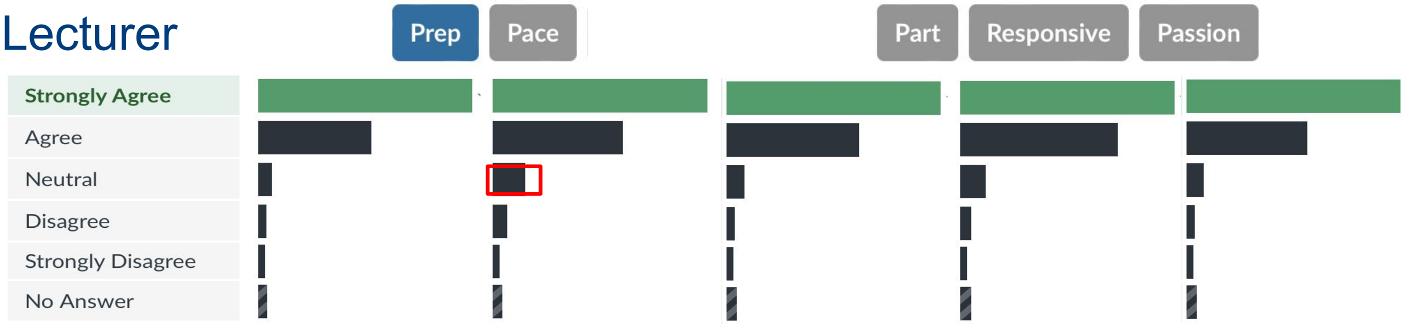
Photo credits: Hans Peters / Anefo @ Wikipedia

Midterm Feedback

Course

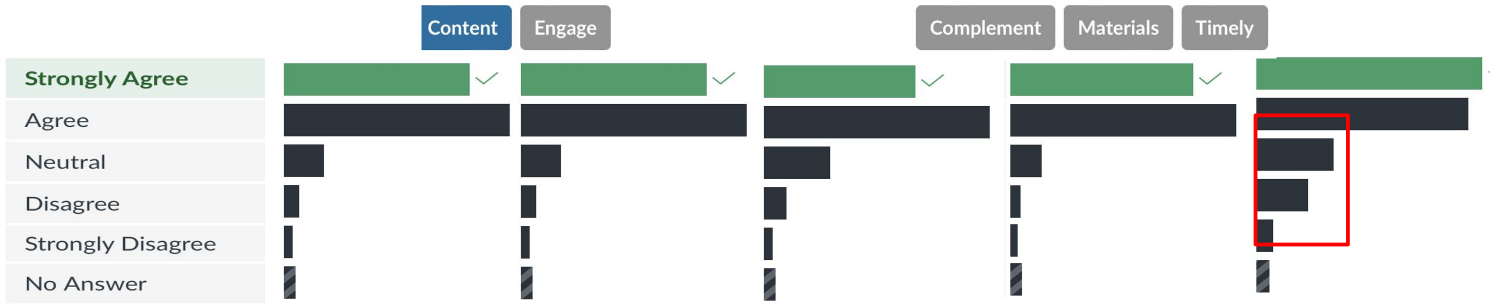


Lecturer

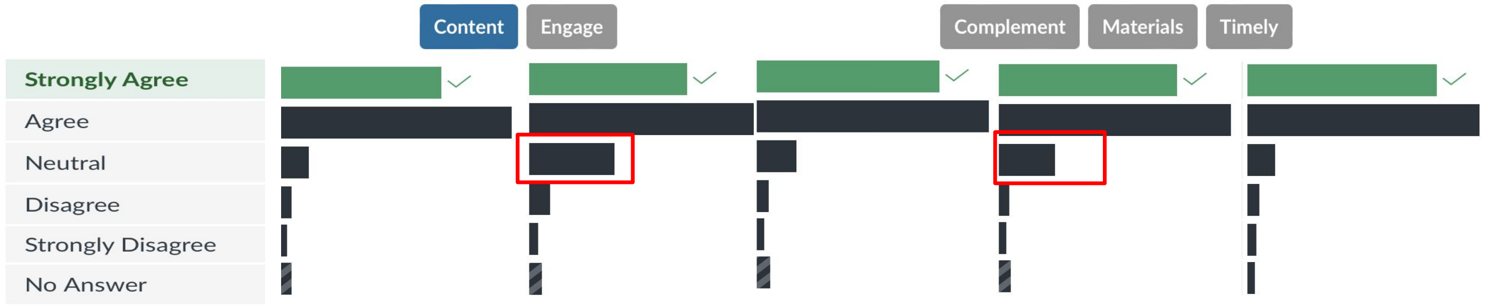


Midterm Feedback: Content

Lecture

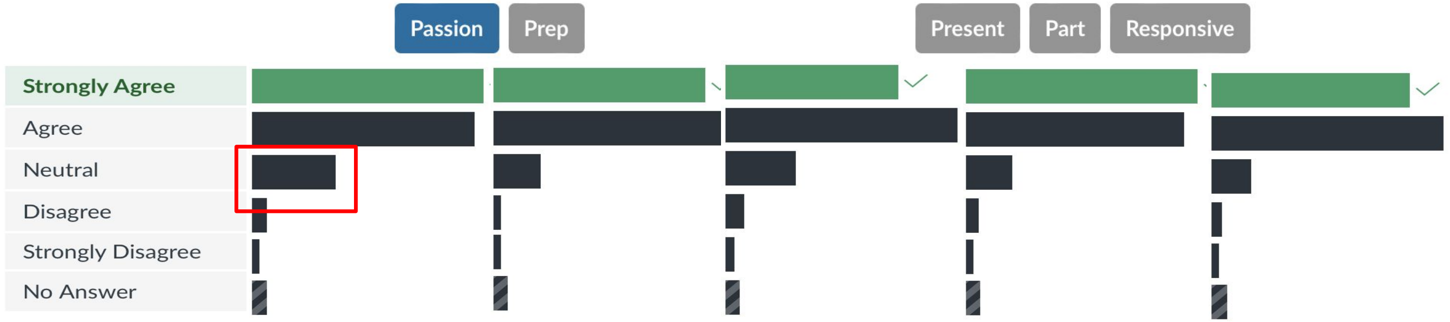


Tutorial

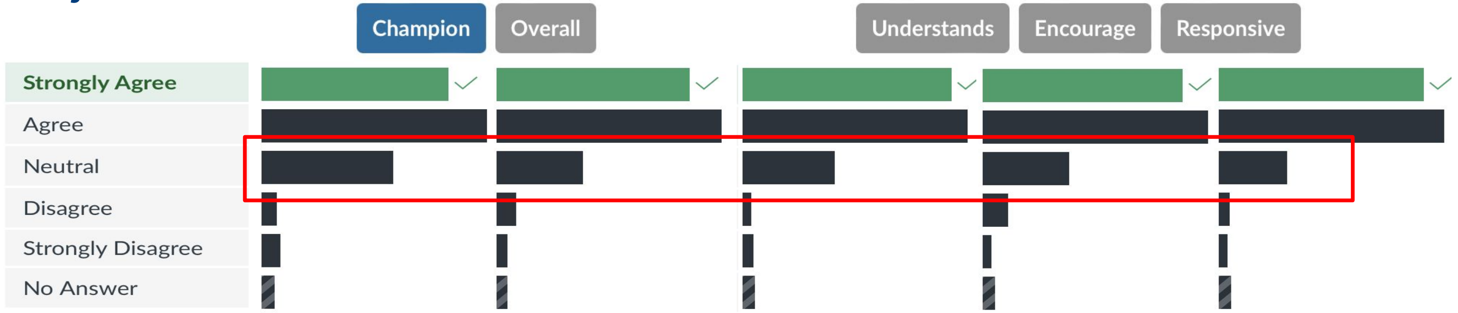


Midterm Feedback: Content

Tutorial Leader



Project Mentor



Improvements

Workload / Assignments

The course could benefit from slightly fewer assignments workload, considering that students also have other subjects to attend to. It's important to focus on the specificity and relevance of the tasks rather than their quantity and complexity, ensuring that each assignment meaningfully contributes to the learning experience.

The workload in this module does seem a bit much. The amount of effort required to do assignment 1 was a lot more than exepcted, and knowing that there were 2 more assignments with the same weightage was quite stressful.

Projects

I feel that the project component is severely lacking in guidance and structure. The only guidance we had was 1. datasets and 2. the rubrics. However, there's no reference point, or specific expectations for the standard we had to hit.

My friends and I found out that if we used the research question in the dataset.pdf, we might get a low grade due to how trivial it is, and we're expected to come up with our own research question, which was not clearly communicated.

Furthermore, having 6, potentially inexperienced people in a group for the project is difficult to manage, and a significant amount of time is spent coordinating the group, rather than the content itself.

Recordings / Length

I refer to recorded lecture material to consolidate and review some concepts I am unable to catch during the actual class. However, the recordings take a long time to be uploaded on Panapto, a week or more. Would appreciate if recordings were uploaded more timely,

The tutorial is too short, instructor often does not go through the entire tutorial sheet. There is also usually no time left at the end of the tutorial slot to go through any additional questions.

The ~3 hrs lecture is too long. Many students, in my honest opinion, do not appreciate the content well enough to be asking relevant questions during lectures.¬†

Outline

- Syntactic Parsing
  - Quick recap: RegEx
  - Context-free grammars
  - Structural Ambiguity
  - Chomsky Normal Form
- **CYK Parsing Algorithm**
  - **Base membership algorithm**
  - Find all parse trees with backtracking
  - Probabilistic parsing
  - Evaluation of parser

CYK Parsing Algorithm

- CYK Parsing Algorithm — basic intuition
  - Given is a context-free grammar G in CNF
  - Assume we have a sentence W comprising n words

There can be multiple rules for different i, but at least one rule for at least one i.

G can generate W $\Longleftrightarrow$ There exists a rule $\boxed{\mathrm{A} \rightarrow \mathrm{B}\ \mathrm{C}}$ in G with

- B can generate $w_1 w_2 w_3 w_4 \dots w_i$
- C can generate $w_{i+1} \dots w_{n-1} w_n$

binary split

$$\underbrace{\overbrace{w_1 w_2 w_3 w_4 \dots w_i}^{B}\underbrace{w_{i+1} \dots w_{n-1} w_n}_{C}}_{A}$$

Note: Appreciate how the "binarization" of rules helps here!

CYK Parsing Algorithm

B_2 C_2

$$\overbrace{w_1 w_2 w_3 w_4 \dots w_i}^{B}\underbrace{w_{i+1} \dots w_{n-1} w_n}_{C}$$

B_1 C_1

- ➜ Recursive nature: G can generate B $\Longleftrightarrow$ There exists a rule $\mathrm{X} \rightarrow \mathrm{Y}\ \mathrm{Z}$ so that X and Y can generate a binary split of $w_1 w_2 w_3 \dots w_i$

$\vdots$

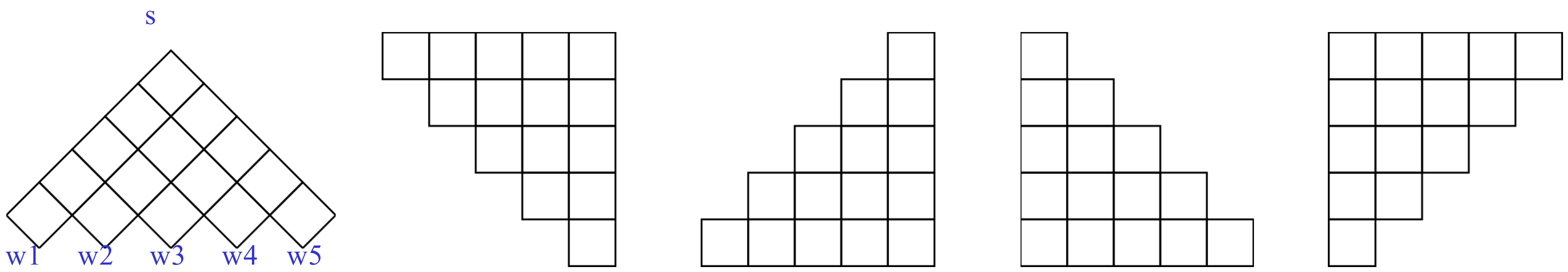
(until we reach individual words; then check the lexicon rules)

- ➜ CYK Parsing Algorithm: Solve problem using Dynamic Programming
  - Find all possible parses for all sequences of size *k* for *k* from 1 to *n*

CYK Parsing Algorithm

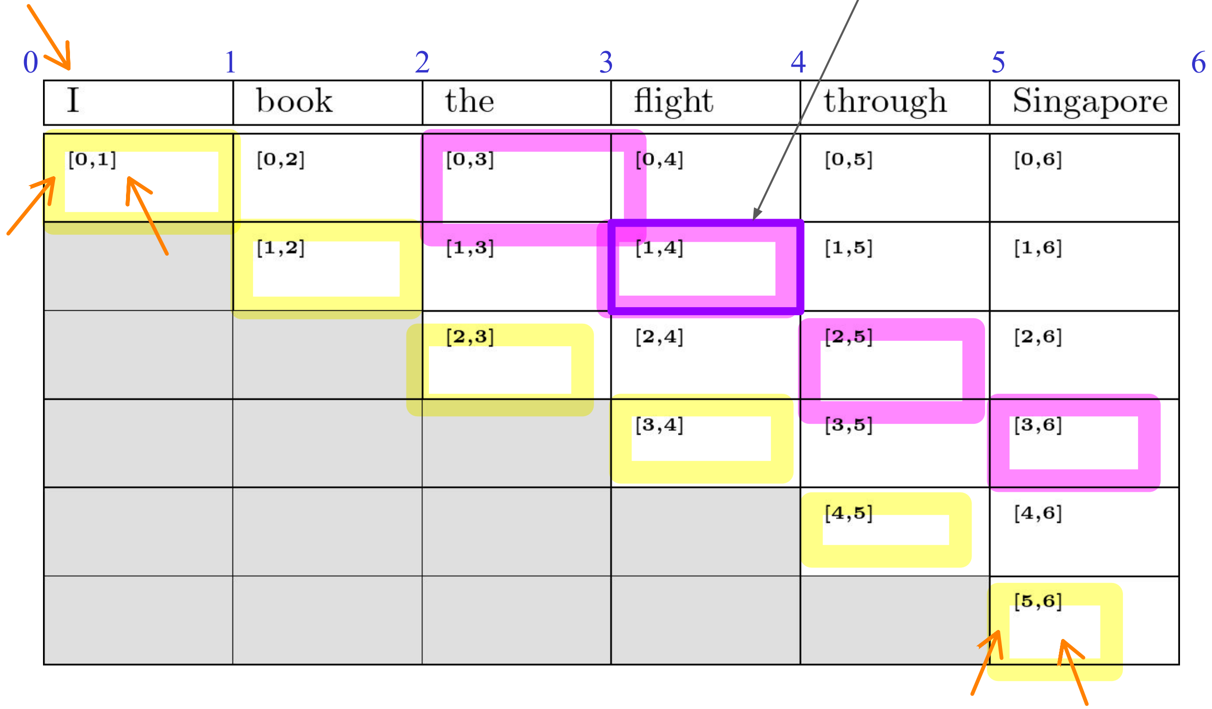
- Dynamic Programming approach
  - Completing the parse table in a bottom-up manner
    (very similar idea as we have seen for calculating the Minimum Edit Distance)
  - Can to handle redundancy when computing the parse trees

- Different ways to visualize parse table
  - Completely identical, just that the indexing of table cells differs



CYK — Parse Table

[1,4] = all possible parses for span "book the flight"



I	book	the	flight	through	Singapore
[0,1]	[0,2]	[0,3]	[0,4]	[0,5]	[0,6]
	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]
		[2,3]	[2,4]	[2,5]	[2,6]
			[3,4]	[3,5]	[3,6]
				[4,5]	[4,6]
					[5,6]

CYK parse table

- $N \times N$ table
  (N = # words in sentence)
- Each cell represents all the possible parses for span [i, j]
- Algorithm: fill table starting with cells for spans of length L = 1 to cells for spans of increasing lengths

L	Cells
1	[0,1], [1,2], [2,3], [3,4], [4,5], [5,6]
2	[0,2], [1,3], [2,4], [3,5], [4,6]
3	[0,3], [1,4], [2,5], [3,6]
4	[0,4], [1,5], [2,6]
5	[0,5], [1,6]
6	[0,6]

In-Lecture Activity (2 mins)

🏃‍♀️🏃‍♂️🏃 Fill in the Lexicon

I	book	the	flight	through	Singapore
[0,1]	[0,2]	[0,3]	[0,4]	[0,5]	[0,6]
	[1,2]	[1,3]	[1,4]	[1,5]	[1,6]
		[2,3]	[2,4]	[2,5]	[2,6]
			[3,4]	[3,5]	[3,6]
				[4,5]	[4,6]
					[5,6]

One of the cells has a pretty interesting value. Which one?

[✓ **A** *book*]

B *flight*

C *through*

D *Singapore*

CYK — Walkthrough

Quick quiz: What does it mean that cell [1,2] contains start symbol S?

I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2]	[0,3]	[0,4]	[0,5]	[0,6]
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4]	[1,5]	[1,6]
		[2,3] Det	[2,4]	[2,5]	[2,6]
			[3,4] Nominal, Noun	[3,5]	[3,6]
				[4,5] Prep	[4,6]
					[5,6] PropNoun, NP

Cells for spans of length L = 1
➜ only need to check lexicon

Fill each cell with the non-terminals that can generate the corresponding word

$$Det \rightarrow the \mid a \mid that \mid this$$
$$Noun \rightarrow book \mid flight \mid meal \mid money$$
$$Verb \rightarrow book \mid include \mid prefer$$
$$Pronoun \rightarrow I \mid she \mid he \mid me$$
$$PropNoun \rightarrow Singapore \mid Frankfurt \mid SIA$$
$$Aux \rightarrow do \mid does \mid did$$
$$Prep \rightarrow from \mid to \mid on \mid near \mid through$$
$$S \rightarrow book \mid include \mid prefer$$
$$VP \rightarrow book \mid include \mid prefer$$
$$NP \rightarrow I \mid she \mid he \mid me$$
$$NP \rightarrow Singapore \mid Frankfurt \mid SIA$$
$$Nominal \rightarrow book \mid flight \mid meal \mid money$$

CYK — Walkthrough

I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4]	[0,5]	[0,6]
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4]	[1,5]	[1,6]
		[2,3] Det	[2,4] NP	[2,5]	[2,6]
			[3,4] Nominal, Noun	[3,5]	[3,6]
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Cells for spans of length L > 1
➜ Check for each binary split if there is a production rule that can generate split

Example: Cell **[0,2]**
➜ only 1 binary split: [0,1] / [1,2]

Check each possible pair of non-terminals of binary split is the RHS of an existing production rule ➜ Yes, add LHS to cell

LHS	RHS
—	Pronoun S
—	Pronoun VP
—	Pronoun Nominal
—	Pronoun Noun
—	Pronoun Verb
—	NP S
S	**NP VP**
—	NP Nominal
—	NP Noun
—	NP Verb

Only this rule exists in our grammar

CYK — Walkthrough

book | the flight

book the | flight

Example: Cell **[1,4]**

- binary split: [1,2] / [2,4]
- binary split: [1,3] / [3,4]

I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4]	[0,5]	[0,6]
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6]
		[2,3] Det	[2,4] NP	[2,5]	[2,6]
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Binary split: [1,2] / [2,4]

LHS	RHS
—	S NP
—	VP NP
—	Nominal NP
—	Noun NP
S, VP, X2	**Verb NP**

3 existing rules with the same RHS

Binary split: [1,3] / [3,4]

LHS	RHS

Empty because [1,3] is empty

Quick quiz: Can you already guess how the parse table indicates that a sentence is valid?

CYK — Walkthrough

I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6]
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6]
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Example: Cell **[2,6]**

- binary split: [2,3] / [3,6]
- binary split: [2,4] / [4,6]
- binary split: [2,5] / [5,6]

Binary split: [2,3] / [3,6]

LHS	RHS
NP	Det Nominal

Binary split: [2,4] / [4,6]

LHS	RHS
—	NP PP

Binary split: [2,5] / [5,6]

LHS	RHS

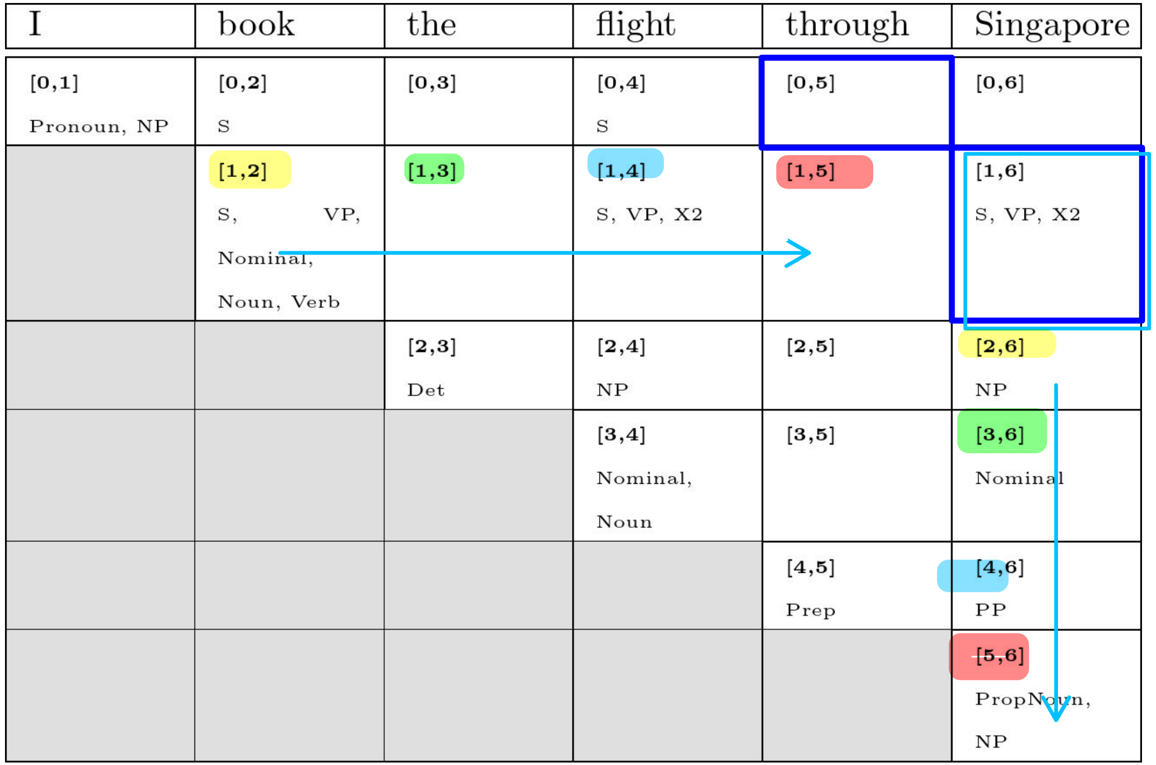
🏃‍♀️🏃🏃 CYK — TraceTogether

Example: Cell **[1,6]**

- binary split: [1,2] / [2,6]
- binary split: [1,3] / [3,6]
- binary split: [1,4] / [4,6]
- binary split: [1,5] / [5,6]

I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6]
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6] S, VP, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

CYK — Walkthrough (Done)



I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6]
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6] S, VP, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Example: Cell **[1,6]**

- binary split: [1,2] / [2,6]
- binary split: [1,3] / [3,6] **(empty!)**
- binary split: [1,4] / [4,6]
- binary split: [1,5] / [5,6] **(empty!)**

Binary split: [1,2] / [2,6]

LHS	RHS
—	S NP
—	VP NP
—	Nominal NP
—	Noun NP
S, VP, X2	**Verb NP**

Binary split: [1,4] / [4,6]

LHS	RHS
—	S PP
S, VP	**VP PP**
S, VP	**X2 PP**

CYK — Walkthrough

I | book the flight through Singapore

I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6] S
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6] S, VP, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Example: Cell **[0,6]**

- binary split: [0,1] / [1,6]
- binary split: [0,2] / [2,6]
- binary split: [0,3] / [3,6] **(empty!)**
- binary split: [0,4] / [4,6]
- binary split: [0,5] / [5,6] **(empty!)**

Binary split: [0,1] / [1,6]

LHS	RHS
—	Pronoun S
—	Pronoun VP
—	Pronoun X2
—	NP S
S	**NP VP**
—	NP X2

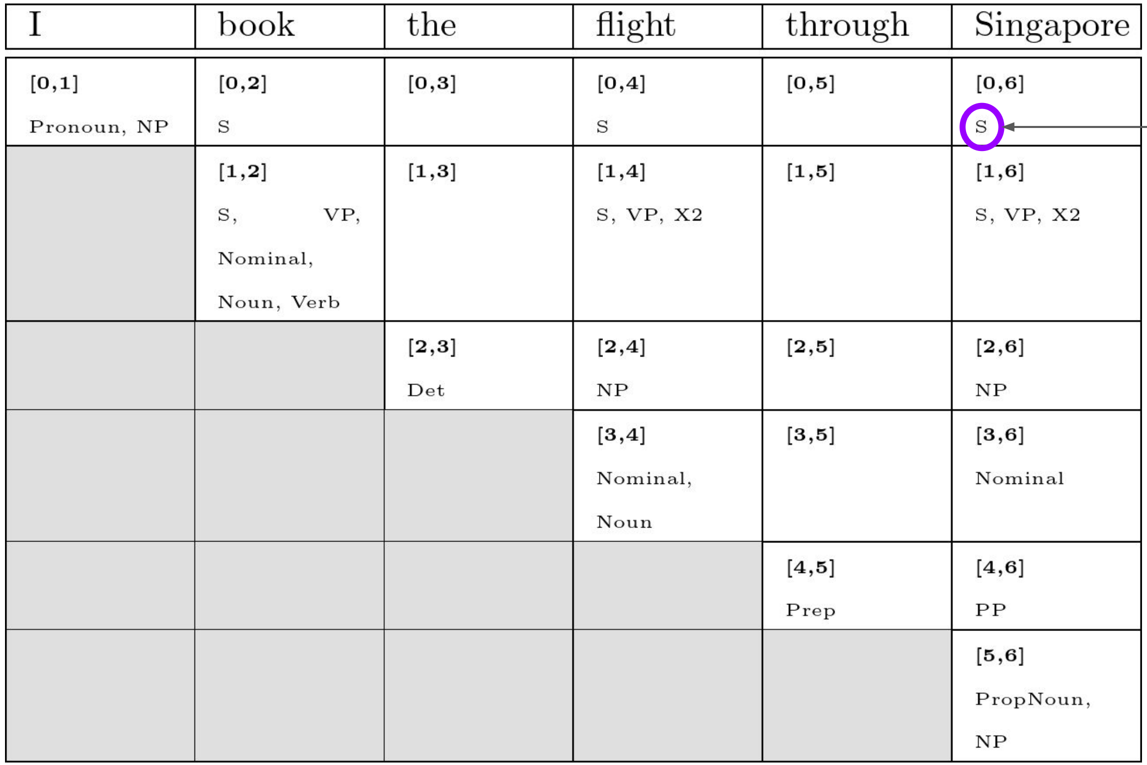
Binary split: [0,2] / [2,6]

LHS	RHS
—	S NP

Binary split: [0,4] / [4,6]

LHS	RHS
—	S PP

CYK — Walkthrough



I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6] S
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6] S, VP, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Our grammar can generate this sentence since the start symbol S is in [0,6]

CYK Parsing Algorithm — Pseudo Code

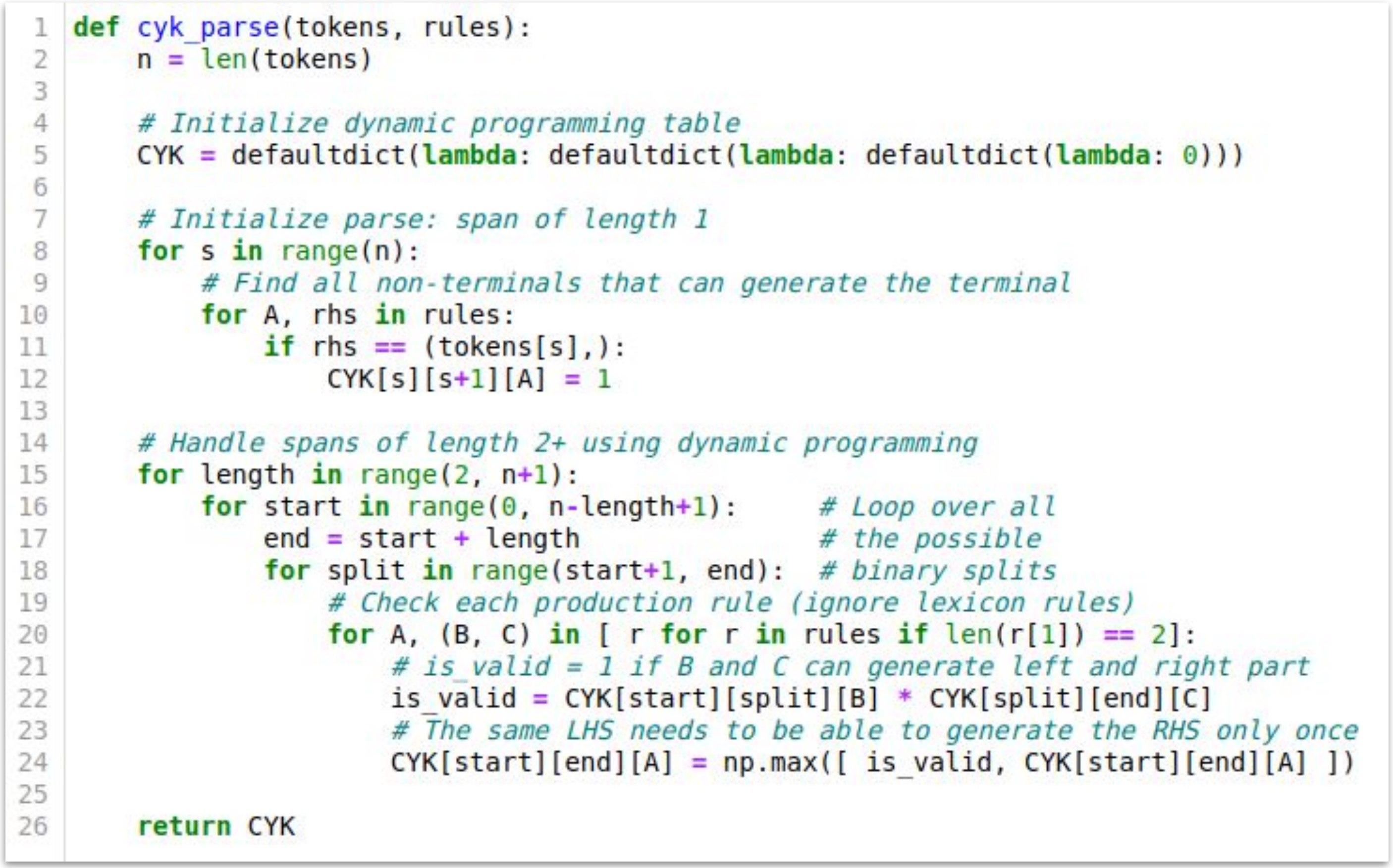
```
function CKY-Parse(words, grammar) returns table

    for j ← from 1 to LENGTH(words) do
        for all { A | A→words[j] ∈ grammar }
            table[j-1, j] ← table[j-1, j] ∪ A

    for j ← from j-2 down to 0 do
        for k←i+1 to j-1 do
            for all { A | A→BC ∈ grammar and B ∈ table[i, k] and C ∈ table[k, j] }
                table[i, j] ← table[i, j] ∪ A
```

Base case: For each terminal (i.e., word), find all terminals that can generate this terminal

Loop over all possible binary splits of spans of size 2, and increase until sentence length

If there is a rule (or more) that can generate the current binary split, add the rule's LHS to the cell of the current span

CYK Parsing Algorithm — Basic Python Implementation



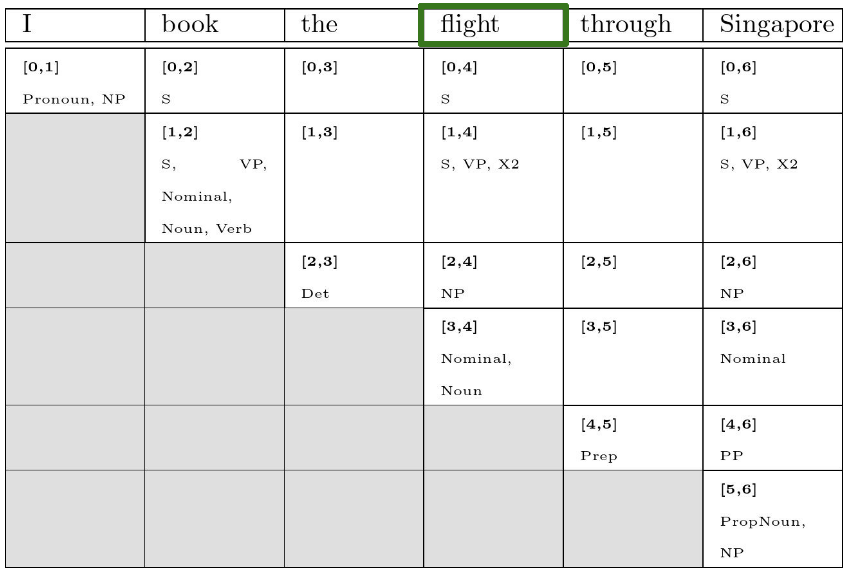
```python
def cyk_parse(tokens, rules):
    n = len(tokens)

    # Initialize dynamic programming table
    CYK = defaultdict(lambda: defaultdict(lambda: defaultdict(lambda: 0)))

    # Initialize parse: span of length 1
    for s in range(n):
        # Find all non-terminals that can generate the terminal
        for A, rhs in rules:
            if rhs == (tokens[s],):
                CYK[s][s+1][A] = 1

    # Handle spans of length 2+ using dynamic programming
    for length in range(2, n+1):
        for start in range(0, n-length+1):     # Loop over all
            end = start + length               # the possible
            for split in range(start+1, end):  # binary splits
                # Check each production rule (ignore lexicon rules)
                for A, (B, C) in [ r for r in rules if len(r[1]) == 2]:
                    # is_valid = 1 if B and C can generate left and right part
                    is_valid = CYK[start][split][B] * CYK[split][end][C]
                    # The same LHS needs to be able to generate the RHS only once
                    CYK[start][end][A] = np.max([ is_valid, CYK[start][end][A] ])

    return CYK
```

CYK — Example: Invalid Parse

I	book	flight	the	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4]	[0,5]	[0,6]
	[1,2] S, VP, Nominal, Noun, Verb	[1,3] Nominal	[1,4]	[1,5]	[1,6]
		[2,3] Nominal, Noun	[2,4]	[2,5]	[2,6]
			[3,4] Det	[3,5]	[3,6]
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

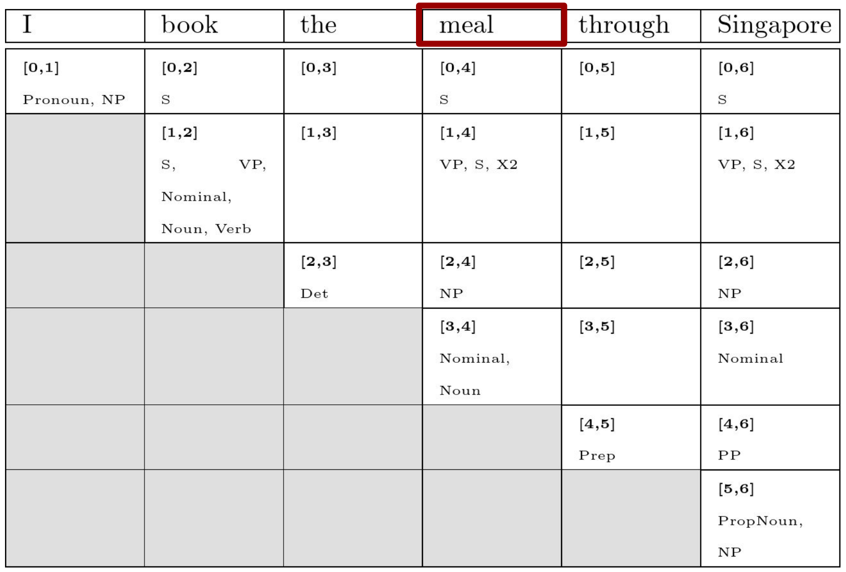
CYK — Syntax vs. Semantic

- Syntactic parsing does not consider semantics
  - Any constituent can be replaced with another constituent of the same type
  - Example below: A noun can be replaced with any other noun



I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6] S
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6] S, VP, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

vs.



I	book	the	meal	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6] S
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] VP, S, X2	[1,5]	[1,6] VP, S, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

CYK Parsing Algorithm — Limitation

- Basic CYK algorithm only solves the membership problem
  - Algorithm only checks if a sentence is a “member” of the language described by the grammar

- What we also want
  - Finding all actual parse trees
    (in case a sentence is valid; otherwise the result is empty)
  - Identifying the best parse tree(s)
    (which requires a definition for what we mean by “best”)

➜ Good news: Only rather minor extension to base algorithm required

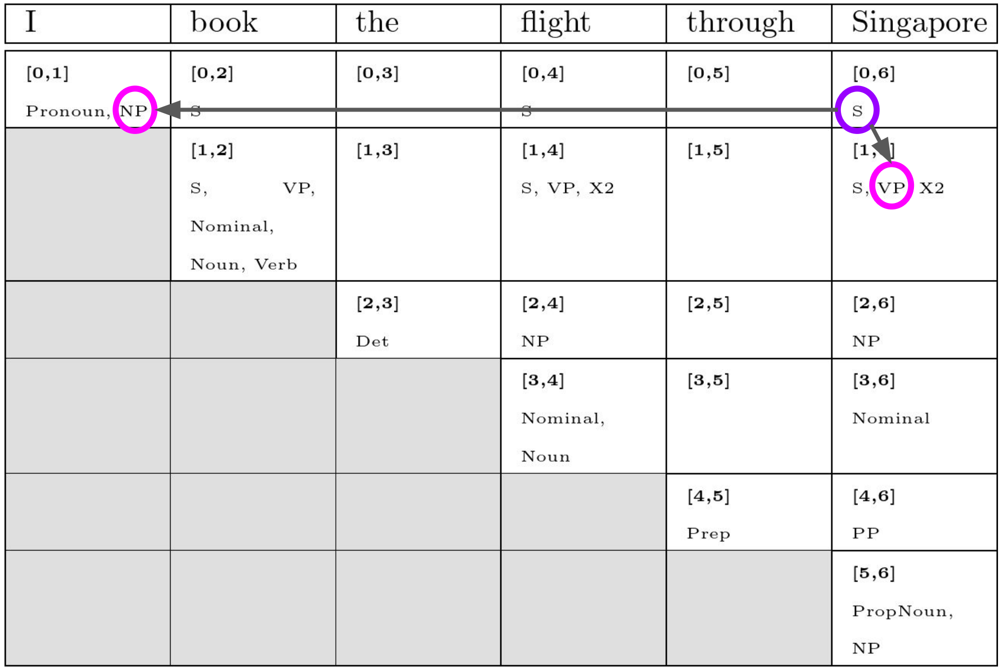
Outline

- Syntactic Parsing
  - Quick recap: RegEx
  - Context-free grammars
  - Structural Ambiguity
  - Chomsky Normal Form

- **CYK Parsing Algorithm**
  - Base membership algorithm
  - **Find all parse trees with backtracking**
  - Probabilistic parsing
  - Evaluation of parser

CYK — Get all Parse Trees (Derivations)

- Basic Idea: Keep track of backtrace
  - Remember which 2 cells matched an existing production rule



I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6] S
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6] S, VP, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Binary split: [0,1] / [1,6]

LHS	RHS
—	Pronoun S
—	Pronoun VP
—	Pronoun X2
—	NP S
S	**NP VP**
—	NP X2

CYK — Get all Parse Trees (Derivations)

- Recall: Structural Ambiguity
  - In general, different production rules might match

I	book	the	flight	through	Singapore
[0,1] Pronoun, NP	[0,2] S	[0,3]	[0,4] S	[0,5]	[0,6] S
	[1,2] S, VP, Nominal, Noun, Verb	[1,3]	[1,4] S, VP, X2	[1,5]	[1,6] S, VP, X2
		[2,3] Det	[2,4] NP	[2,5]	[2,6] NP
			[3,4] Nominal, Noun	[3,5]	[3,6] Nominal
				[4,5] Prep	[4,6] PP
					[5,6] PropNoun, NP

Binary split: [1,2] / [2,6]

LHS	RHS
—	S NP
—	VP NP
—	Nominal NP
—	Noun NP
S, VP, X2	Verb NP

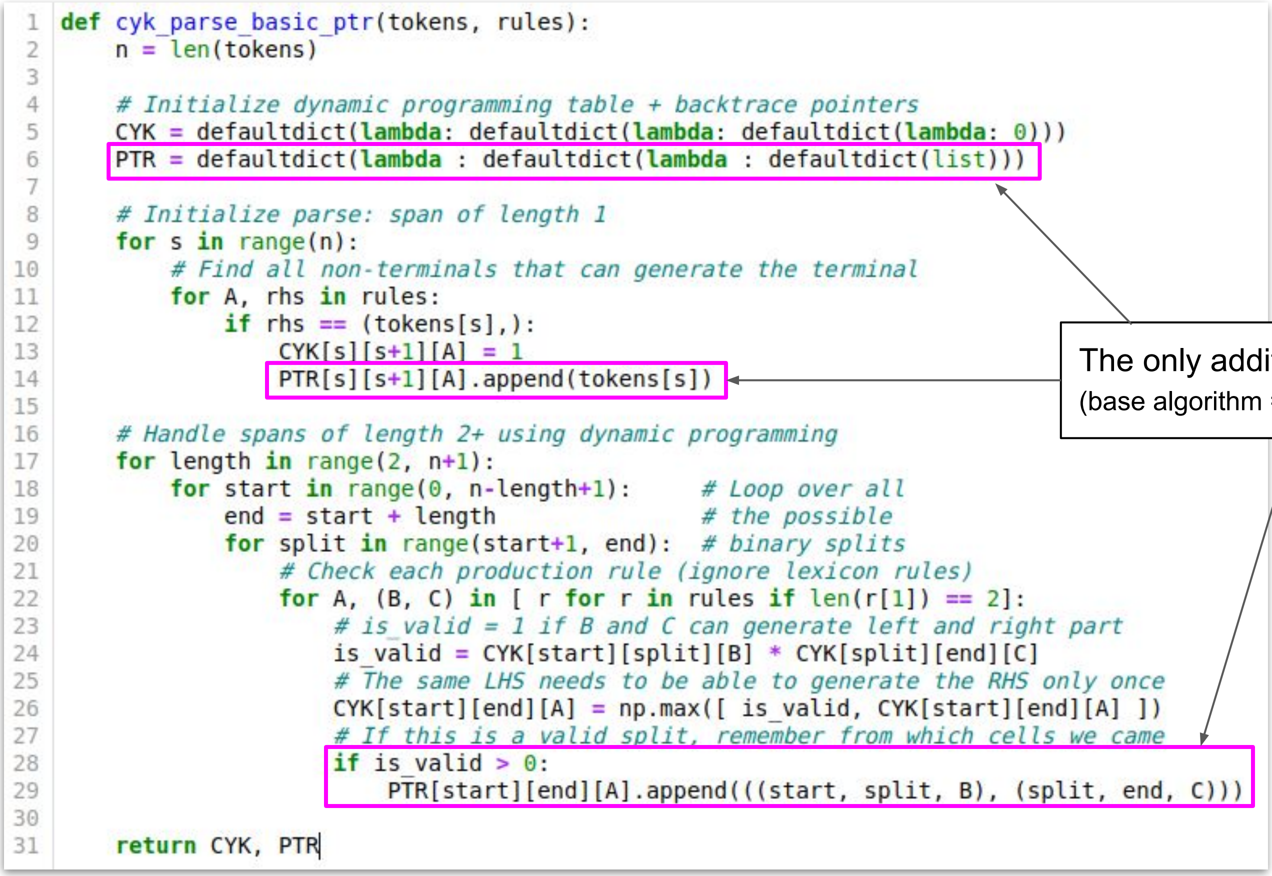
Binary split: [1,4] / [4,6]

LHS	RHS
—	S PP
S, VP	VP PP
S, VP	X2 PP

CYK — Get all Parse Trees (Pseudo Code)

```
function CKY-Parse(words, grammer) returns table, pointer

    for j ← from 1 to LENGTH(words) do
        for all { A | A→words[j] ∈ grammar }
            table[j-1, j] ← table[j-1, j] ∪ A
            pointer[j-1, j, A] ← pointer[j-1, j, A] ∪ words[j]
                                                  }
    for j ← from j-2 down to 0 do
        for k←i+1 to j-1 do
            for all { A | A→BC ∈ grammar and B ∈ table[i, k] and C ∈ table[k, j] }
                table[i, j] ← table[i, j] ∪ A
                pointer[i, j, A] ← pointer[i, j, A] ∪ ((i, k, B), (k, j, C))
```

CYK — Get all Parse Trees (Python)



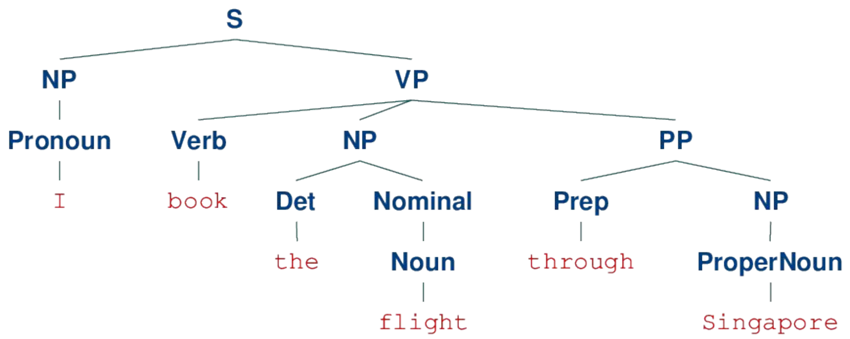
```python
def cyk_parse_basic_ptr(tokens, rules):
    n = len(tokens)

    # Initialize dynamic programming table + backtrace pointers
    CYK = defaultdict(lambda: defaultdict(lambda: defaultdict(lambda: 0)))
    PTR = defaultdict(lambda : defaultdict(lambda : defaultdict(list)))

    # Initialize parse: span of length 1
    for s in range(n):
        # Find all non-terminals that can generate the terminal
        for A, rhs in rules:
            if rhs == (tokens[s],):
                CYK[s][s+1][A] = 1
                PTR[s][s+1][A].append(tokens[s])

    # Handle spans of length 2+ using dynamic programming
    for length in range(2, n+1):
        for start in range(0, n-length+1):      # Loop over all
            end = start + length                # the possible
            for split in range(start+1, end):   # binary splits
                # Check each production rule (ignore lexicon rules)
                for A, (B, C) in [ r for r in rules if len(r[1]) == 2]:
                    # is_valid = 1 if B and C can generate left and right part
                    is_valid = CYK[start][split][B] * CYK[split][end][C]
                    # The same LHS needs to be able to generate the RHS only once
                    CYK[start][end][A] = np.max([ is_valid, CYK[start][end][A] ])
                    # If this is a valid split, remember from which cells we came
                    if is_valid > 0:
                        PTR[start][end][A].append(((start, split, B), (split, end, C)))

    return CYK, PTR
```

The only additions to the base algorithm
(base algorithm = CYK for membership problem)

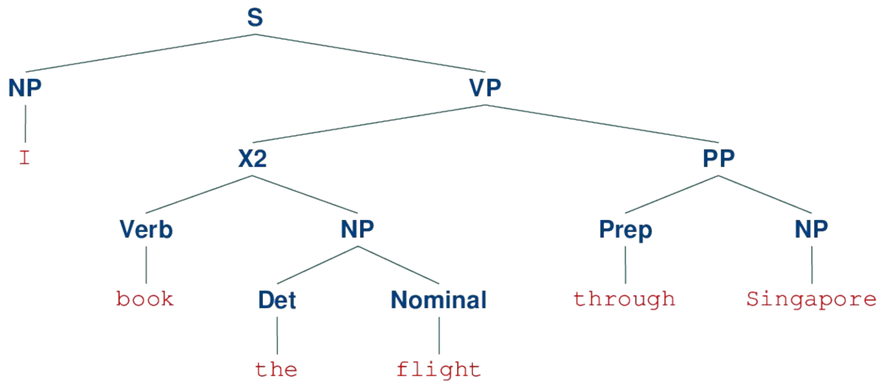
Parse Trees: CFG vs. CNF

- Converting a CFG into a CNF affects resulting parse trees
  - CFG parse trees can be recovered from CNF parse trees
    (easy for newly split n-ary rules; a bit more tricky for unary rules)
  - Straightforward extension of CYK algorithm to support unary rules directly
    (doesn't affect runtime complexity, but roughly doubles the required lines code)

Parse tree using original CFG



Parse tree using CNF (converted from CFG)

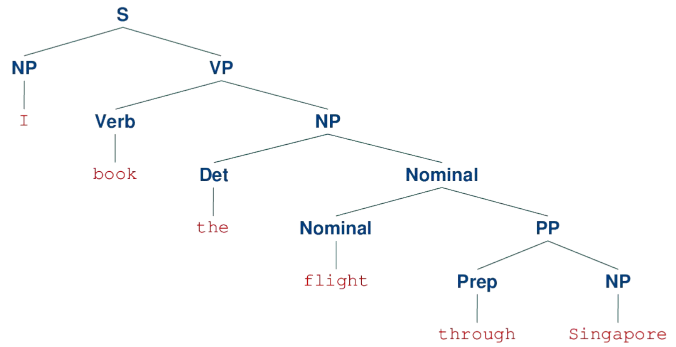


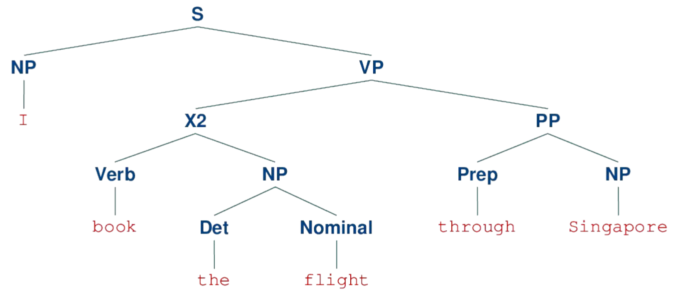
CYK — Parse Trees

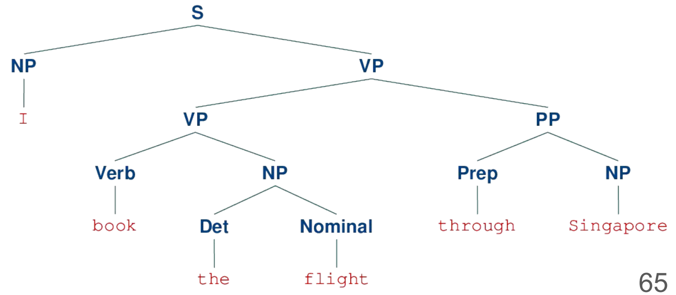
- Parse tree for example

  "I book the flight through Singapore"

- Observation
  - Multiple valid parses
  - Which is the best one?







🏃‍♀️🏃🏃‍♂️ Looking for ~~a few good~~ the best parses

- Question:

  🤔 So how would you find the best parse?

  🧐 Propose a solution.

- Post your answer to Canvas > Discussions > [In-Lecture Interaction] L1

  (Help like other classmate's responses too! 👍)

Outline

- Syntactic Parsing
  - Quick recap: RegEx
  - Context-free grammars
  - Structural Ambiguity
  - Chomsky Normal Form

- **CYK Parsing Algorithm**
  - Base membership algorithm
  - Find all parse trees with backtracking
  - **Probabilistic parsing**
  - Evaluation of parser

Statistical Parsing

- Resolve structural ambiguity by choosing the most probable parse
  - Best parse = parse with the highest probability
  - Question: Where to get such probabilities from?

➜ Probabilistic Context-Free Grammar (PCFG)

- Same as CFG, but each rule is associated with a probability
- Probabilities are learned from an annotated dataset

Given a parse tree T for a sentence S comprised of rules: $P(T,S) = \prod_{i}^{n} P(A \rightarrow \alpha) = \prod_{i}^{n} P(\alpha|A)$

CFG — Formal Definition

- A CFG is a 4 tuple $\langle N, \Sigma, R, S \rangle$
  - N — set of non-terminal symbols
  - Σ — set of terminal symbols
  - R — set of rules

Allowed format for all rules

$A \rightarrow \alpha\ [p]$ with $A \in N,\ \alpha \in (N \cup \Sigma)$

$p = P(\alpha|A)$

  - S — start symbol

Example

$N = \{\text{NP},\ \text{VP},\ \text{Det},\ \text{Noun},\ \text{Verb}\}$

$\Sigma = \{a,\ the,\ man,\ meal,\ flight,\ saw,\ booked\}$

$$\begin{aligned} \text{S} &\rightarrow \text{NP VP } [0.4] \\ \text{NP} &\rightarrow \text{Det Noun } [0.5] \\ \text{VP} &\rightarrow \text{Verb NP } [0.2] \\ \text{Det} &\rightarrow a\ [0.3] \mid the\ [0.4] \\ &\dots \end{aligned}$$

Example CFG → Example PCFG

$$\sum = 1 \left\{ \begin{array}{r} S \rightarrow NP\ VP\ [0.8] \\ S \rightarrow Aux\ NP\ VP\ [0.1] \\ S \rightarrow VP\ [0.1] \end{array} \right.$$

$$\sum = 1 \left\{ \begin{array}{r} NP \rightarrow Pronoun\ [0.2] \\ NP \rightarrow ProperNoun\ [0.2] \\ NP \rightarrow Det\ Nominal\ [0.6] \end{array} \right.$$

$$\sum = 1 \left\{ \begin{array}{r} Nominal \rightarrow Noun\ [0.3] \\ Nominal \rightarrow Nominal\ Noun\ [0.2] \\ Nominal \rightarrow Nominal\ PP\ [0.5] \end{array} \right.$$

$$\sum = 1 \left\{ \begin{array}{r} VP \rightarrow Verb\ [0.2] \\ VP \rightarrow Verb\ NP\ [0.4] \\ VP \rightarrow Verb\ NP\ PP\ [0.1] \\ VP \rightarrow Verb\ PP\ [0.1] \\ VP \rightarrow VP\ PP\ [0.2] \\ PP \rightarrow Prep\ NP\ [1.0] \end{array} \right.$$

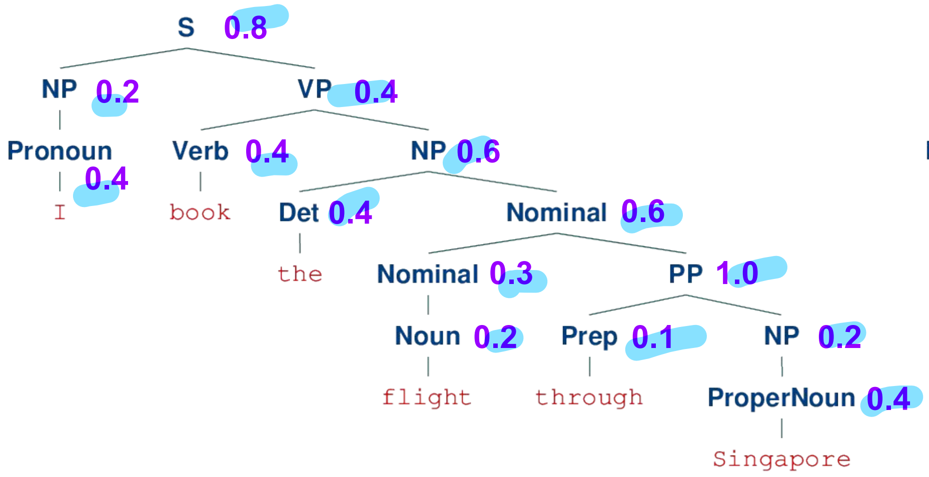
$$\begin{array}{r} Det \rightarrow \\ Noun \rightarrow \\ Verb \rightarrow \\ Pronoun \rightarrow \\ ProperNoun \rightarrow \\ Aux \rightarrow \\ Prep \rightarrow \end{array} \underbrace{\begin{array}{l} the\ [0.4] \mid a\ [0.3] \mid that\ [0.2] \mid this\ [0.1] \\ book\ [0.2] \mid flight\ [0.2] \mid meal\ [0.3] \mid money\ [0.3] \\ book\ [0.4] \mid include\ [0.3] \mid prefer\ [0.3] \\ I\ [0.4] \mid she\ [0.2] \mid he\ [0.2] \mid me\ [0.2] \\ Singapore\ [0.4] \mid Frankfurt\ [0.4] \mid SIA\ [0.2] \\ do\ [0.5] \mid does\ [0.2] \mid did\ [0.3] \\ from\ [0.2] \mid to\ [0.4] \mid on\ [0.2] \mid near\ [0.1] \mid through\ [0.1] \end{array}}_{\sum = 1 \text{ for all right-hand sides}}$$

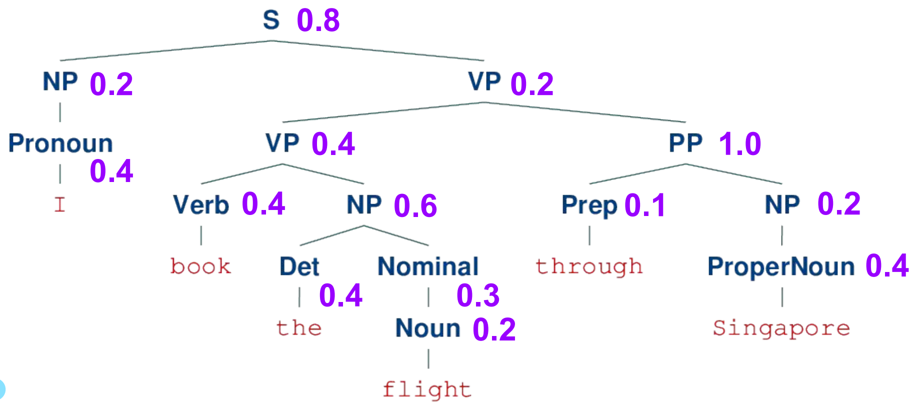
Requirement for valid probabilities:

$$\sum_{\alpha} P(A \rightarrow \alpha) = \sum_{\alpha} P(\alpha | A) = 1$$

PCFG — Probability of a Parse Tree

- Probability of parse tree = product of probabilities of all rules
  - In practice, sum up log probabilities to avoid arithmetic underflow





$$P(T,S) = \prod_{i}^{n} P(A \rightarrow \alpha) = 0.00000071$$

$$P(T,S) = \prod_{i}^{n} P(A \rightarrow \alpha) = 0.00000024$$

PCFG — Calculating the Probability of a Rule

- Calculating $P(A \rightarrow \alpha)$ using Maximum Likelihood Estimation
  - Requires annotated dataset of parse trees

Number of occurrences of rule $P(A \rightarrow \alpha)$ in the dataset

$$P(A \rightarrow \alpha) = P(\alpha|A) = \frac{Count(A \rightarrow \alpha)}{Count(A)}$$

Number of occurrences of rules in the dataset with A as the LHS

PCFG — Converting to CNF

(1) Dividing n-ary rules by introducing new non-terminals

$$\begin{array}{l} S \rightarrow NP\ VP\ [0.8] \\ S \rightarrow Aux\ NP\ VP\ [0.1] \\ \ldots \end{array} \quad \Longrightarrow \quad \begin{array}{l} S \rightarrow NP\ VP\ [0.8] \\ S \rightarrow X1\ VP\ [0.1] \\ X1 \rightarrow Aux\ NP\ [1.0] \\ \ldots \end{array}$$

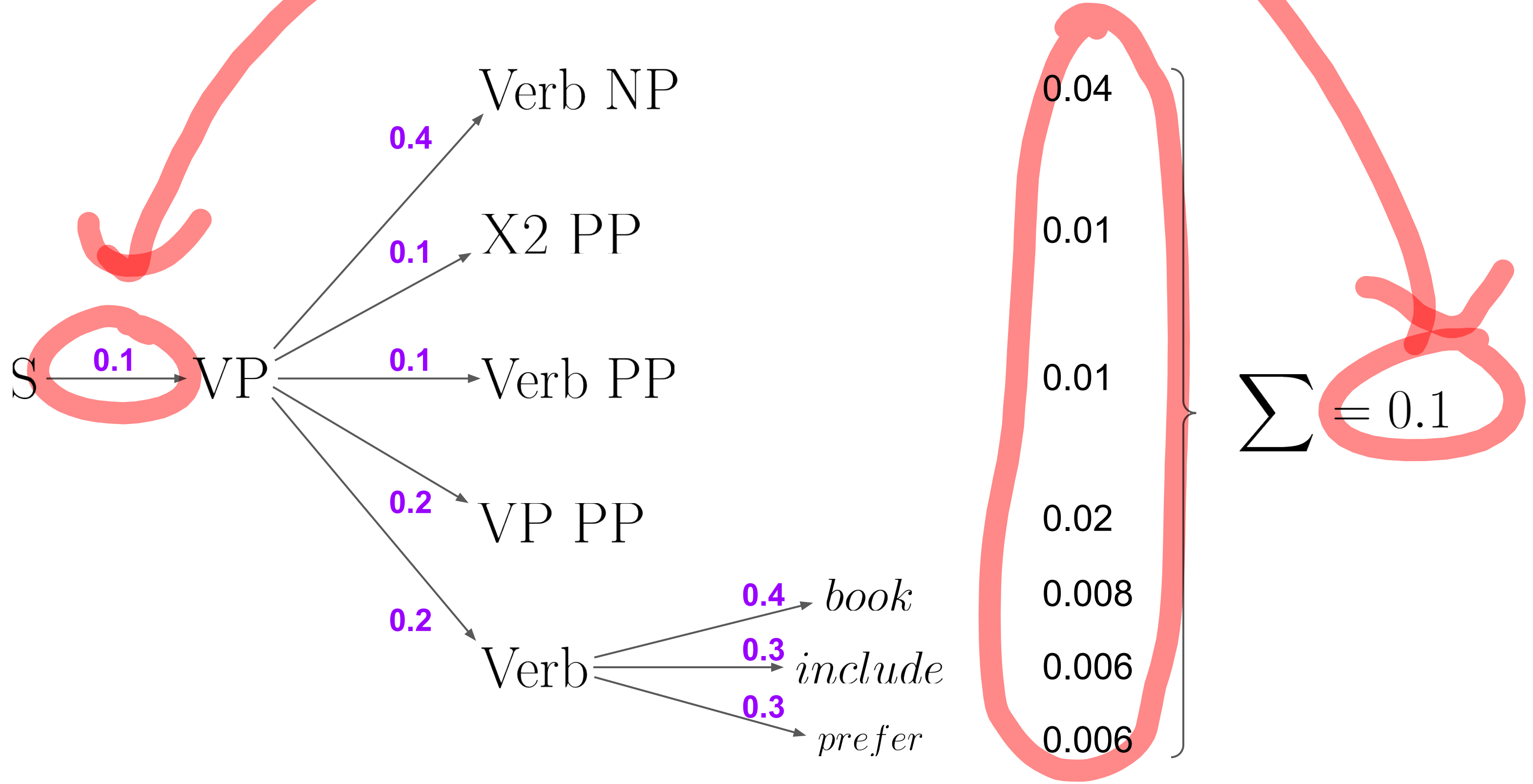
(2) Recursive removal of unary rules

$$\begin{array}{l} S \rightarrow NP\ VP\ [0.8] \\ S \rightarrow Aux\ NP\ VP\ [0.1] \\ S \rightarrow VP\ [0.1] \\ \ldots \end{array} \quad \Longrightarrow \quad \begin{array}{l} S \rightarrow NP\ VP\ [0.8] \\ S \rightarrow X1\ VP\ [0.1] \\ X1 \rightarrow Aux\ NP\ [1.0] \\ \boxed{\begin{array}{l} S \rightarrow Verb\ NP\ [0.04] \\ S \rightarrow X2\ PP\ [0.01] \\ S \rightarrow Verb\ PP\ [0.01] \\ S \rightarrow VP\ PP\ [0.02] \\ S \rightarrow \mathit{book}\ [0.008] \mid \mathit{include}\ [0.006] \mid \mathit{prefer}\ [0.006] \end{array}} \end{array}$$

How to compute these probabilities?

PCFG — Converting to CNF

- Multiply probabilities along the paths



PCFG — Converting to CNF

$$\begin{aligned}
\text{S} &\rightarrow \text{NP VP } [0.8] \\
\text{S} &\rightarrow \text{X1 VP } [0.1] \\
\text{X1} &\rightarrow \text{Aux NP } [1.0] \\
\text{S} &\rightarrow \text{Verb NP } [0.04] \\
\text{S} &\rightarrow \text{X2 PP } [0.01] \\
\text{S} &\rightarrow \text{Verb PP } [0.01] \\
\text{S} &\rightarrow \text{VP PP } [0.02] \\
\text{NP} &\rightarrow \text{Det Nominal } [0.6] \\
\text{Nominal} &\rightarrow \text{Nominal Noun } [0.2] \\
\text{Nominal} &\rightarrow \text{Nominal PP } [0.5] \\
\text{VP} &\rightarrow \text{Verb NP } [0.2] \\
\text{VP} &\rightarrow \text{X2 PP } [0.1] \\
\text{X2} &\rightarrow \text{Verb NP } [1.0] \\
\text{VP} &\rightarrow \text{Verb PP } [0.1] \\
\text{VP} &\rightarrow \text{VP PP } [0.2] \\
\text{PP} &\rightarrow \text{Prep NP } [1.0]
\end{aligned}$$

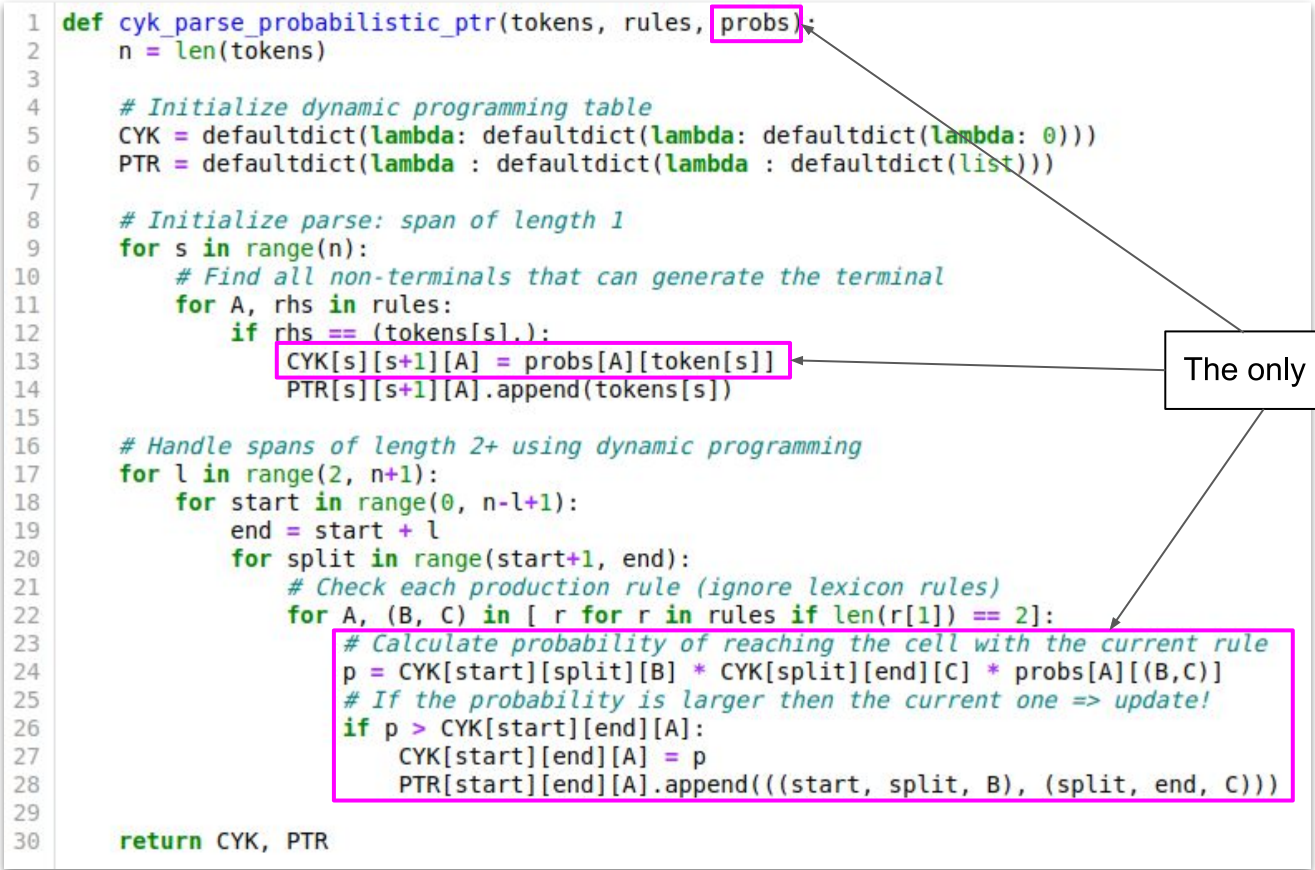
$$\begin{aligned}
\text{Det} &\rightarrow the\ [0.4] \mid a\ [0.3] \mid that\ [0.2] \mid this\ [0.1] \\
\text{Noun} &\rightarrow book\ [0.2] \mid flight\ [0.2] \mid meal\ [0.3] \mid money\ [0.3] \\
\text{Verb} &\rightarrow book\ [0.4] \mid include\ [0.3] \mid prefer\ [0.3] \\
\text{Pronoun} &\rightarrow I\ [0.4] \mid she\ [0.2] \mid he\ [0.2] \mid me\ [0.2] \\
\text{PropNoun} &\rightarrow Singapore\ [0.4] \mid Frankfurt\ [0.4] \mid SIA\ [0.2] \\
\text{Aux} &\rightarrow do\ [0.5] \mid does\ [0.2] \mid did\ [0.3] \\
\text{Prep} &\rightarrow from\ [0.2] \mid to\ [0.4] \mid on\ [0.2] \mid near\ [0.1] \mid through\ [0.1] \\
\text{S} &\rightarrow book\ [0.008] \mid include\ [0.006] \mid prefer\ [0.006] \\
\text{VP} &\rightarrow book\ [0.08] \mid include\ [0.06] \mid prefer\ [0.06] \\
\text{NP} &\rightarrow I\ [0.08] \mid she\ [0.04] \mid he\ [0.04] \mid me\ [0.04] \\
\text{NP} &\rightarrow Singapore\ [0.08] \mid Frankfurt\ [0.08] \mid SIA\ [0.04] \\
\text{Nominal} &\rightarrow book\ [0.06] \mid flight\ [0.06] \mid meal\ [0.09] \mid money\ [0.09]
\end{aligned}$$

pen & paper calculations…I hope the numbers add up :)

CYK — Get Best Parse Tree (Pseudo Code)

```
function CKY-Parse(words, grammar, probs) returns table, pointer

    for j ← from 1 to LENGTH(words) do                    compile partial parses for n-grams of increasing size
        for all { A | A→words[j] ∈ grammar }
            table[j-1, j, A] ← probs[A→words[j]]          compute probability
            pointer[j-1, j, A] ← pointer[j-1, j, A] ∪ words[j]

    for j ← from j-2 down to 0 do
        for k←i+1 to j-1 do
            for all { A | A→BC ∈ grammar and B ∈ table[i, k] and C ∈ table[k, j] }
                p ← table[i, k, B] * table[k, j, C] * probs[A→BC]
                if p > table[i, j, A] do
                    table[i, j, A] ← p
                    pointer[i, j, A] ← pointer[i, j, A] ∪ ((i, k, B), (k, j, C))
```

CYK — Get Best Parse Tree (Python)



```python
def cyk_parse_probabilistic_ptr(tokens, rules, probs):
    n = len(tokens)

    # Initialize dynamic programming table
    CYK = defaultdict(lambda: defaultdict(lambda: defaultdict(lambda: 0)))
    PTR = defaultdict(lambda : defaultdict(lambda : defaultdict(list)))

    # Initialize parse: span of length 1
    for s in range(n):
        # Find all non-terminals that can generate the terminal
        for A, rhs in rules:
            if rhs == (tokens[s],):
                CYK[s][s+1][A] = probs[A][token[s]]
                PTR[s][s+1][A].append(tokens[s])

    # Handle spans of length 2+ using dynamic programming
    for l in range(2, n+1):
        for start in range(0, n-l+1):
            end = start + l
            for split in range(start+1, end):
                # Check each production rule (ignore lexicon rules)
                for A, (B, C) in [ r for r in rules if len(r[1]) == 2]:
                    # Calculate probability of reaching the cell with the current rule
                    p = CYK[start][split][B] * CYK[split][end][C] * probs[A][(B,C)]
                    # If the probability is larger then the current one => update!
                    if p > CYK[start][end][A]:
                        CYK[start][end][A] = p
                        PTR[start][end][A].append(((start, split, B), (split, end, C)))

    return CYK, PTR
```

The only changes to the algorithm

Outline

- Syntactic Parsing
  - Quick recap: RegEx
  - Context-free grammars
  - Structural Ambiguity
  - Chomsky Normal Form

- **CYK Parsing Algorithm**
  - Base membership algorithm
  - Find all parse trees with backtracking
  - Probabilistic parsing
  - **Evaluation of parser**

Note: This is not the only way to do it.

Evaluation of Parse Trees

- Important: best parse $\nRightarrow$ correct parse
  - Best parse = parse with the highest probability
  - Correct parse = parse that matches the gold-standard solution

- How evaluate parse trees?
  - Represent each parse tree as a set of tuples $\{\langle l_1, i_1, j_1 \rangle, \langle l_2, i_2, j_2 \rangle, ..., \langle l_n, i_n, j_n \rangle\}$

    l_k is the non-terminal labeling the kth phrase

    i_k is the index of the first word in the kth word in the phrase

    j_k is the index of the last word in the kth word in the phrase

A

B

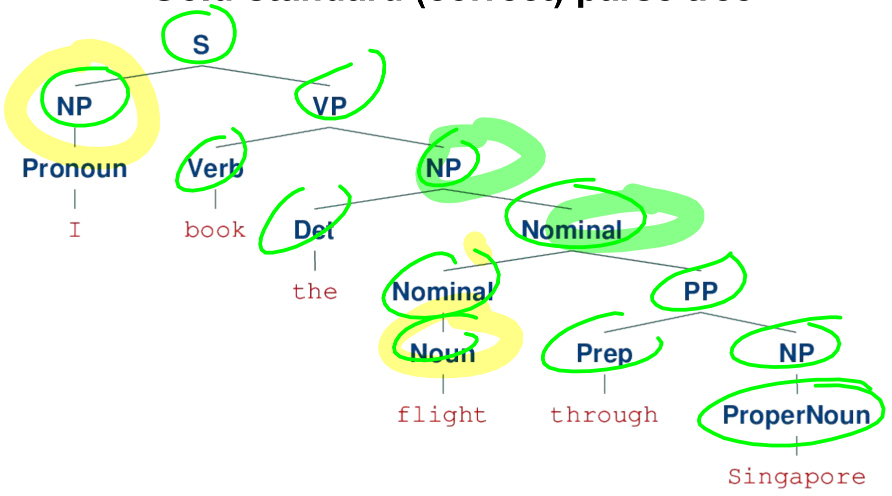
C

$A \rightarrow B\ C$

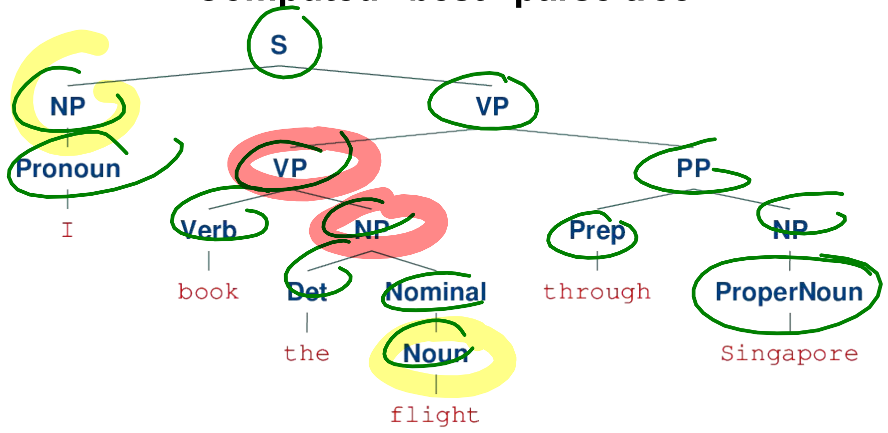
➜ Use representations of computed parse and gold standard parse to estimate precision, recall and F1

Evaluation of Parse Trees — Example

Gold-standard (correct) parse tree



Computed “best” parse tree



Tuples only present in correct tree

$\langle \text{NP}, 3, 6 \rangle$

$\langle \text{Nominal}, 4, 6 \rangle$

Tuples resent in both trees

$\langle \text{NP}, 1, 1 \rangle$ $\langle \text{Pronoun}, 1, 1 \rangle$ $\langle \text{VP}, 2, 2 \rangle$ $\langle \text{Verb}, 2, 2 \rangle$

$\langle \text{Det}, 3, 3 \rangle$ $\langle \text{Nominal}, 4, 4 \rangle$ $\langle \text{Noun}, 4, 4 \rangle$ $\langle \text{Prep}, 5, 5 \rangle$

$\langle \text{ProperNoun}, 6, 6 \rangle$ $\langle \text{PP}, 5, 6 \rangle$ $\langle \text{NP}, 6, 6 \rangle$

Tuples only present in computed tree

$\langle \text{VP}, 2, 4 \rangle$

$\langle \text{NP}, 3, 4 \rangle$

Evaluation of Parse Trees — Example

Tuples only present in correct tree	Tuples resent in both trees	Tuples only present in computed tree
$\langle \text{NP}, 3, 6\rangle$	$\langle \text{NP}, 1, 1\rangle$ $\langle \text{Pronoun}, 1, 1\rangle$ $\langle \text{VP}, 2, 2\rangle$ $\langle \text{Verb}, 2, 2\rangle$	$\langle \text{VP}, 2, 4\rangle$
$\langle \text{Nominal}, 4, 6\rangle$	$\langle \text{Det}, 3, 3\rangle$ $\langle \text{Nominal}, 4, 4\rangle$ $\langle \text{Noun}, 4, 4\rangle$ $\langle \text{Prep}, 5, 5\rangle$	$\langle \text{NP}, 3, 4\rangle$
	$\langle \text{ProperNoun}, 6, 6\rangle$ $\langle \text{PP}, 5, 6\rangle$ $\langle \text{NP}, 6, 6\rangle$	

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{11}{11+2} = 0.85$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{11}{11+2} = 0.85$$

$$\text{f1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = 0.85$$

TP = # tuples in both trees

FP = # tuples only in the computed tree

FN = # tuples only in the correct tree

Summary

- Recursive nature of natural language
  - Natural language allows for nested structure
  - Basic building block: **constituents**
  - Most common way to capture constituency ➜ **context-free grammars (CFGs)**

- Syntactic parsing
  - Membership: check if a sentence can be generated by a grammar
  - Identification of <u>all</u> possible parse trees for a sentence
  - Identification of <u>best</u> parse tree for a sentence ➜ **Probabilistic CFGs**

$\}$ **CYK Parsing Algorithm**

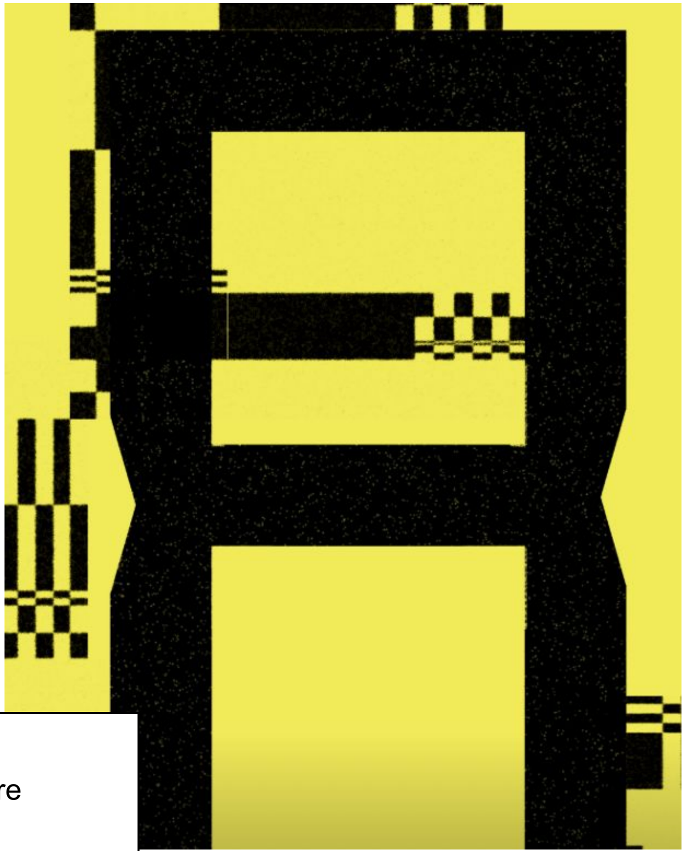
Outlook for Next Week: Transformers

Photo credit: Hasbro and Paramouht Pictures

Pre-Lecture Activity for Next Week

Read *8 Google Employees Invented Modern AI. Here's the Inside Story*
(Wired Article)

Apply your own (self-)attention to the article. Quote a sentence of the article you think most or least strikes your attention. Tell us why.



Side notes:

- This task is meant as a warm-up to provide some context for the next lecture
- No worries if you get lost; we will talk about this in the next lecture
- You can just copy-&-paste others' answers but this won't help you learn better

Photo credit: Wired