Automata and Formal Languages (7)

Email:	christian.urban at kcl.ac.uk
Office:	S1.27 (1st floor Strand Building)
Slides:	KEATS (also home work is there)

Two Weeks Ago: CFGs

A context-free grammar (CFG) G consists of:

- a finite set of nonterminal symbols (upper case)
- a finite terminal symbols or tokens (lower case)
- a start symbol (which must be a nonterminal)
- a set of rules

$$A \rightarrow \mathrm{rhs}_1 | \mathrm{rhs}_2 | \ldots$$

where rhs are sequences involving terminals and nonterminals (can also be empty).

Two Weeks Ago: CFGs

A context-free grammar (CFG) G consists of:

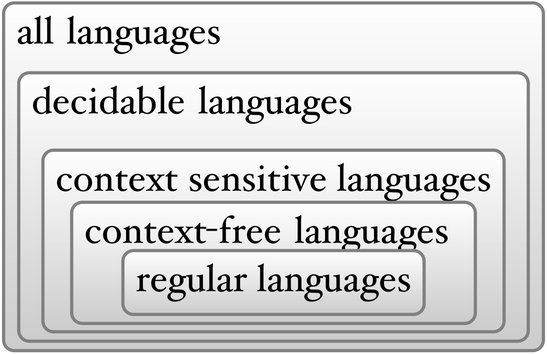
- a finite set of nonterminal symbols (upper case)
- a finite terminal symbols or tokens (lower case)
- a start symbol (which must be a nonterminal)
- a set of rules

$$A \rightarrow \text{rhs}_1 | \text{rhs}_2 | \ldots$$

where rhs are sequences involving terminals and nonterminals (can also be empty).

Hierarchy of Languages

Recall that languages are sets of strings.



Arithmetic Expressions

A grammar for arithmetic expressions and numbers:

$$\begin{array}{lcl} \boldsymbol{E} & \rightarrow & \boldsymbol{E} \cdot + \cdot \boldsymbol{E} \mid \boldsymbol{E} \cdot * \cdot \boldsymbol{E} \mid (\cdot \boldsymbol{E} \cdot) \mid \boldsymbol{N} \\ \boldsymbol{N} & \rightarrow & \boldsymbol{N} \cdot \boldsymbol{N} \mid 0 \mid 1 \mid \ldots \mid 9 \end{array}$$

Unfortunately it is left-recursive (and ambiguous).

A problem for recursive descent parsers (e.g. parser combinators).

Arithmetic Expressions

A grammar for arithmetic expressions and numbers:

$$\begin{array}{lcl} \boldsymbol{E} & \rightarrow & \boldsymbol{E} \cdot + \cdot \boldsymbol{E} \mid \boldsymbol{E} \cdot * \cdot \boldsymbol{E} \mid (\cdot \boldsymbol{E} \cdot) \mid \boldsymbol{N} \\ \boldsymbol{N} & \rightarrow & \boldsymbol{N} \cdot \boldsymbol{N} \mid 0 \mid 1 \mid \ldots \mid 9 \end{array}$$

Unfortunately it is left-recursive (and ambiguous).

A problem for recursive descent parsers (e.g. parser combinators).

Numbers

$$N \rightarrow N \cdot N \mid 0 \mid 1 \mid \ldots \mid 9$$

A non-left-recursive, non-ambiguous grammar for numbers:

$$N \rightarrow 0 \cdot N \mid 1 \cdot N \mid \ldots \mid 0 \mid 1 \mid \ldots \mid 9$$

Operator Precedences

To disambiguate

$$E \quad \rightarrow \quad E \cdot + \cdot E \mid E \cdot * \cdot E \mid (\cdot E \cdot) \mid N$$

Decide on how many precedence levels, say

highest for $()$, medium for $*$, lowest for $+$

$$\begin{array}{lcl} E_{low} & \rightarrow & E_{med} \cdot + \cdot E_{low} \mid E_{med} \\ E_{med} & \rightarrow & E_{hi} \cdot * \cdot E_{med} \mid E_{hi} \\ E_{hi} & \rightarrow & (\cdot E_{low} \cdot) \mid N \end{array}$$

Operator Precedences

To disambiguate

$$E \rightarrow E \cdot + \cdot E \mid E \cdot * \cdot E \mid (\cdot E \cdot) \mid N$$

Decide on how many precedence levels, say

highest for $()$, medium for $*$, lowest for $+$

$$\begin{array}{lcl} E_{low} & \rightarrow & E_{med} \cdot + \cdot E_{low} \mid E_{med} \\ E_{med} & \rightarrow & E_{hi} \cdot * \cdot E_{med} \mid E_{hi} \\ E_{hi} & \rightarrow & (\cdot E_{low} \cdot) \mid N \end{array}$$

What happens with $1 + 3 + 4$?

Removing Left-Recursion

The rule for numbers is directly left-recursive:

$$N \rightarrow N \cdot N \mid 0 \mid 1 \quad (\ldots)$$

Translate

$$\begin{array}{lll} N & \rightarrow & N \cdot \alpha \\ & \mid & \beta \end{array} \quad \Rightarrow \quad \begin{array}{lll} N & \rightarrow & \beta \cdot N' \\ N' & \rightarrow & \alpha \cdot N' \\ & \mid & \epsilon \end{array}$$

Removing Left-Recursion

The rule for numbers is directly left-recursive:

$$N \quad \rightarrow \quad N \cdot N \mid 0 \mid 1 \qquad (\ldots)$$

Translate

$$\begin{array}{lll} N & \rightarrow & N \cdot \alpha \\ & \mid & \beta \end{array} \quad \Rightarrow \quad \begin{array}{lll} N & \rightarrow & \beta \cdot N' \\ N' & \rightarrow & \alpha \cdot N' \\ & \mid & \epsilon \end{array}$$

Which means

$$\begin{array}{lll} N & \rightarrow & 0 \cdot N' \mid 1 \cdot N' \\ N' & \rightarrow & N \cdot N' \mid \epsilon \end{array}$$

Chomsky Normal Form

All rules must be of the form

$$A \rightarrow a$$

or

$$A \rightarrow B \cdot C$$

No rule can contain ϵ.

ϵ-Removal

1. If $A \rightarrow \alpha \cdot B \cdot \beta$ and $B \rightarrow \epsilon$ are in the grammar, then add $A \rightarrow \alpha \cdot \beta$ (iterate if necessary).
2. Throw out all $B \rightarrow \epsilon$.

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \\ N' &\rightarrow N \cdot N' \mid \epsilon \end{aligned}$$

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\ N' &\rightarrow N \cdot N' \mid N \mid \epsilon \end{aligned}$$

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\ N' &\rightarrow N \cdot N' \mid N \end{aligned}$$

ϵ-Removal

1. If $A \rightarrow \alpha \cdot B \cdot \beta$ and $B \rightarrow \epsilon$ are in the grammar, then add $A \rightarrow \alpha \cdot \beta$ (iterate if necessary).
2. Throw out all $B \rightarrow \epsilon$.

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \\ N' &\rightarrow N \cdot N' \mid \epsilon \end{aligned}$$

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\ N' &\rightarrow N \cdot N' \mid N \mid \epsilon \end{aligned}$$

$$\begin{aligned} N &\rightarrow 0 \cdot N' \mid 1 \cdot N' \mid 0 \mid 1 \\ N' &\rightarrow N \cdot N' \mid N \end{aligned}$$

$$N \rightarrow 0 \cdot N \mid 1 \cdot N \mid 0 \mid 1$$

CYK Algorithm

If grammar is in Chomsky normalform ...

$$
\begin{array}{lcl}
S & \rightarrow & N \cdot P \\
P & \rightarrow & V \cdot N \\
N & \rightarrow & N \cdot N \\
N & \rightarrow & \texttt{students} \mid \texttt{Jeff} \mid \texttt{geometry} \mid \texttt{trains} \\
V & \rightarrow & \texttt{trains}
\end{array}
$$

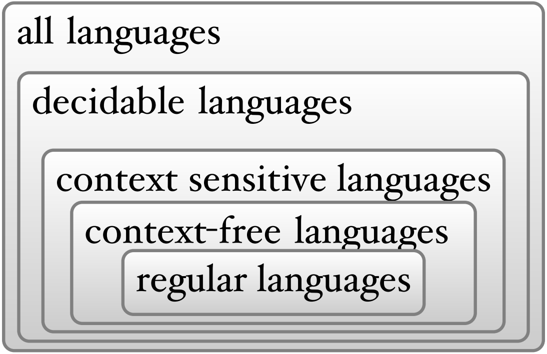
`Jeff trains geometry students`

CYK Algorithm

- fastest possible algorithm for recognition problem
- runtime is $O(n^3)$
- grammars need to be transferred into CNF

Hierarchy of Languages

Recall that languages are sets of strings.



Context Sensitive Grms

$$
\begin{array}{lcl}
S & \Rightarrow & bSAA \mid \epsilon \\
A & \Rightarrow & a \\
bA & \Rightarrow & Ab
\end{array}
$$

Context Sensitive Grms

$$
\begin{array}{lcl}
S & \Rightarrow & bSAA \mid \epsilon \\
A & \Rightarrow & a \\
bA & \Rightarrow & Ab
\end{array}
$$

$$S \Rightarrow \ldots \Rightarrow^{?} \text{"}ababaa\text{"}$$

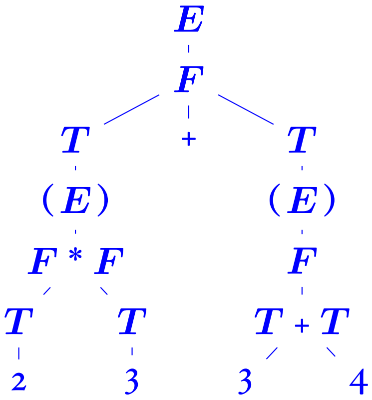
Parse Trees

$$
\begin{array}{lcl}
\mathit{Stmt} & \rightarrow & \texttt{skip} \\
 & | & \mathit{Id}\ \texttt{:=}\ \mathit{AExp} \\
 & | & \texttt{if}\ \mathit{BExp}\ \texttt{then}\ \mathit{Block}\ \texttt{else}\ \mathit{Block} \\
 & | & \texttt{while}\ \mathit{BExp}\ \texttt{do}\ \mathit{Block} \\
 & | & \texttt{read}\ \mathit{Id} \\
 & | & \texttt{write}\ \mathit{Id} \\
 & | & \texttt{write}\ \mathit{String} \\
\mathit{Stmts} & \rightarrow & \mathit{Stmt}\ \texttt{;}\ \mathit{Stmts} \\
 & | & \mathit{Stmt} \\
\mathit{Block} & \rightarrow & \{\ \mathit{Stmts}\ \} \\
 & | & \mathit{Stmt} \\
\mathit{AExp} & \rightarrow & \ldots
\end{array}
$$

Parse Trees

$$
\begin{array}{lcl}
E & \rightarrow & F \mid F \cdot * \cdot F \\
F & \rightarrow & T \mid T \cdot + \cdot T \mid T \cdot - \cdot T \\
T & \rightarrow & num\_token \mid (\cdot E \cdot)
\end{array}
$$

(2*3)+(3+4)



Ambiguous Grammars

A CFG is ambiguous if there is a string that has at least parse trees.

$$
\begin{array}{lcl}
E & \rightarrow & num\_token \\
E & \rightarrow & E \cdot + \cdot E \\
E & \rightarrow & E \cdot - \cdot E \\
E & \rightarrow & E \cdot * \cdot E \\
E & \rightarrow & (\cdot E \cdot)
\end{array}
$$

1 + 2 * 3 + 4

Dangling Else

Another ambiguous grammar:

$$\begin{array}{rcl} E & \rightarrow & \text{if } E \text{ then } E \\ & | & \text{if } E \text{ then } E \text{ else } E \\ & | & \text{id} \end{array}$$

```
if a then if x then y else c
```

A CFG Derivation

1. Begin with a string with only the start symbol S
2. Replace any non-terminal X in the string by the right-hand side of some production $X \rightarrow \text{rhs}$
3. Repeat 2 until there are no non-terminals

$$S \rightarrow \ldots \rightarrow \ldots \rightarrow \ldots \rightarrow \ldots$$