CS310 : Automata Theory 2019

Lecture 14: Context-free grammars

Instructor: Ashutosh Gupta

IITB, India

Compile date: 2019-02-05

Beyond regular languages

We know many languages are not regular.

More importantly, many useful and well-structured languages are not regular.

Example 14.1

The language L_{bal} of balanced parentheses is not regular.

- $() \in L_{bal}$
- $(())() \in L_{bal}$
- $(()() \notin L_{bal}$

Let us build a theory to handle languages like this!

Example: production rules

Example 14.2

L_{bal} can be inductively described as follows

base case:

ϵ is in L_{bal}

induction step:

There are two ways to create larger words from L_{bal}

1. *If $w \in L_{bal}$, then (w) is also in L_{bal}*
2. *If $w_1, w_2 \in L_{bal}$, then then $w_1 w_2$ is also in L_{bal}*

We will write the above rules formally as follows

1. $B \to \epsilon$
2. $B \to (B)$
3. $B \to BB$

} *Context-free grammar*

Where B represents a set of words produced by the above rules.

Elements of context-free grammars(CFG)

1. Terminals: The symbols that occur in words.
   e. g., {'(', ')' }

2. Nonterminals: Symbols that represent a class of words.
   e. g., $\{B\}$

3. Production rules A production rules consists the following three parts
   - Left hand side containing a single nonterminal
   - Production symbol $\rightarrow$
   - Right hand side containing a word over terminals and nonterminals

   e. g., $B \rightarrow (B)$

4. Start symbol One of the nonterminal is the start symbol.
   e. g., B is the start symbol for the *CFG* for L_{bal}.

More interesting context-free grammar

The language is clearly not regular.

Example 14.3

Consider the language of arithmetic expressions over binary numbers, e. g.,

- $10 + 1$
- $(10 + 101) \times 100$

$T = \{+, \times, 0, 1, (,)\}$ *are the terminals.*

We will have two nonterminals $\{B, E\}$

- B *is a nonterminal for binary numbers*
- E *is a nonterminal for the arithmetic expressions*

E *is the start symbol.*

More interesting context-free grammar II

The following are the production rules.

Actually B is a regular language $1(0+1)^*$

1. $B \to 1$
2. $B \to B0$
3. $B \to B1$

} rules for producing B

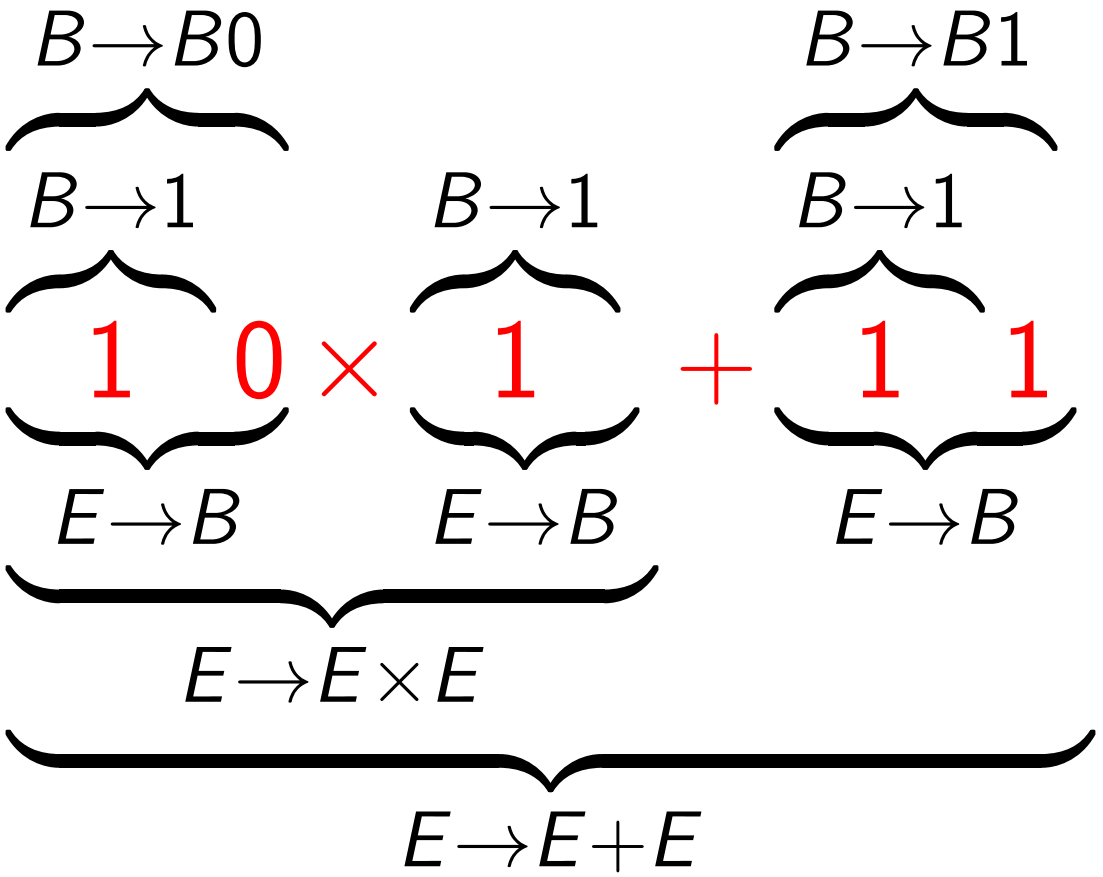
4. $E \to B$
5. $E \to E + E$
6. $E \to E \times E$
7. $E \to (E)$

} rules for producing E

From grammar to words

We obtain words of the language by constructing words by recursively applying production rules.

Example 14.4



We will formalize the above illustration!

Writing compact CFG

There is a compact way of writing the CFG production rules.

We can club the rules for a nonterminal together and separate right hand sides using "|".

Example 14.5

For our example,

- $B \to 1 \mid B0 \mid B1$
- $E \to B \mid E + E \mid E \times E \mid (E)$

Context-free grammar (CFG)

Definition 14.1

A Context-free grammar(CFG) G is a four-tuple

$$(N, T, P, S)$$

where

- *N is the set of nonterminals*
- *T is the set of terminals,*
- *P is the set of production rules, and*
- *$S \in T$ is the start symbol.*

Recall we defined DFAs first and REs later. Breaking from the pattern, we have defined CFGs first and will define corresponding automata later.

Example: context-free grammar

Exercise 14.1

Formally, the grammar for arithmetic expressions is

$$G_{arith} = (\underbrace{\{E, B\}}_{\text{Nonterminals}}, \underbrace{\{+, \times, 0, 1, (,)\}}_{\text{Terminals}}, P, \underbrace{E}_{\text{Start symbol}})$$

where, P contains the following production rules.

1. $B \to 1$
2. $B \to B0$
3. $B \to B1$
4. $E \to B$
5. $E \to E + E$
6. $E \to E \times E$
7. $E \to (E)$

Derivations using context-free grammar

The natural interpretation of a grammar $G = (N, T, P, S)$ is a single step derivation.

Let us suppose we have a word $\alpha A \beta$, where $\alpha, \beta \in (N \cup T)^*$ and $A \in N$.

Let $A \to \gamma \in P$.

We say we derive $\alpha\gamma\beta$ from $\alpha A \beta$ after applying rule $A \to \gamma$.

Definition 14.2

Formally, a single step derivation relation $\overset{G}{\Rightarrow}$ is written as follows.

$$\alpha A \beta \overset{G}{\Rightarrow} \alpha\gamma\beta$$

If grammar is obvious from the context we may only write $\Rightarrow$.

Mutiple derivations

We define application of $\overset{G}{\Rightarrow}$ zero or more states as follows.

Definition 14.3

$\alpha \overset{G*}{\Longrightarrow} \beta$ *if there is a sequence of words* $\gamma_1, \ldots \gamma_n$ *such that*

- $\alpha = \gamma_1$,
- $\beta = \gamma_n$, *and*
- $\gamma_{i-1} \overset{G}{\Rightarrow} \gamma_i$ *for each* $i \in 2..n$.

Example: grammar derivations

Example 14.6

Consider again the following arithmetic expression grammar.

1. $B \to 1$
2. $B \to B0$
3. $B \to B1$
4. $E \to B$
5. $E \to E + E$
6. $E \to E \times E$
7. $E \to (E)$

Let us see the derivation $10 \times 1 + 11$ *in the above grammar.*

$$E \Rightarrow E + E \Rightarrow E + B \Rightarrow E \times E + B \Rightarrow E \times B + B \Rightarrow E \times 1 + B$$

$$\Rightarrow B \times 1 + B \Rightarrow B0 \times 1 + B \Rightarrow B0 \times 1 + B1 \Rightarrow 10 \times 1 + B1 \Rightarrow 10 \times 1 + 11$$

Too many choices for derivation

We can expand any nonterminal in α.

The choices seem to be like nondeterminism of NFA.

We can limit the derivation moves and have predictable sequence of derivations, i.e., making derivations deterministic.

Leftmost derivation

Always expand leftmost nonterminal. We denote such derivation by $\overset{lm}{\Rightarrow}$

Example 14.7

Consider again the following arithmetic expression grammar.

1. $B \to 1$
2. $B \to B0$
3. $B \to B1$
4. $E \to B$
5. $E \to E + E$
6. $E \to E \times E$
7. $E \to (E)$

Let us see the leftmost derivation of $10 \times 1 + 11$ *in the above grammar.*

$$E \overset{lm}{\Rightarrow} E + E \overset{lm}{\Rightarrow} E \times E + E \overset{lm}{\Rightarrow} B \times E + E \overset{lm}{\Rightarrow} B0 \times E + E \overset{lm}{\Rightarrow} 10 \times E + E$$
$$\overset{lm}{\Rightarrow} 10 \times B + E \overset{lm}{\Rightarrow} 10 \times 1 + E \overset{lm}{\Rightarrow} 10 \times 1 + B \overset{lm}{\Rightarrow} 10 \times 1 + B1 \overset{lm}{\Rightarrow} 10 \times 1 + 11$$

Rightmost derivation

Always expand rightmost nonterminal. We denote such derivation by $\overset{rm,G}{\Longrightarrow}$

Example 14.8

Consider again the following arithmetic expression grammar.

1. $B \to 1$
2. $B \to B0$
3. $B \to B1$
4. $E \to B$
5. $E \to E + E$
6. $E \to E \times E$
7. $E \to (E)$

Let us see the rightmost derivation of $10 \times 1 + 11$ *in the above grammar.*

$$E \overset{rm}{\Longrightarrow} E + E \overset{rm}{\Longrightarrow} E + B \overset{rm}{\Longrightarrow} E + B1 \overset{rm}{\Longrightarrow} E + 11 \overset{rm}{\Longrightarrow} E \times E + 11$$
$$\overset{rm}{\Longrightarrow} E \times B + 11 \overset{rm}{\Longrightarrow} E \times 1 + 11 \overset{rm}{\Longrightarrow} B \times 1 + 11 \overset{rm}{\Longrightarrow} B0 \times 1 + 11 \overset{rm}{\Longrightarrow} 10 \times 1 + 11$$

What is "context-free" in context-free grammar?

In word $\alpha A \beta \in (N + T)^*$, α and β has no influence on expansion of $A \in N$.

In other words,

an application of a production rule for A is

independent

of the context of A.

Language of a grammar

Definition 14.4

Let $G = (N, T, P, S)$ be a CFG. The language of G, denoted $L(G)$, is the set of terminal strings that have derivations from the start symbol.

$$L(G) \triangleq \{w \in T^* | S \overset{G*}{\Longrightarrow} w\}$$

If a language L is the language of some context-free grammar, then we say L is a context-free language.

Words for nonterminals

Definition 14.5

For a grammar $G = (N, T, P, S)$, $A \in N$, *and* $\alpha \in (N \cup T)^*$,
if $A \overset{G*}{\Longrightarrow} \alpha$ *we say* α *is a word of* A.

Example 14.9

Consider the following derivations in the arithmetic expressions grammar

$$B \Rightarrow B1 \Rightarrow 11$$

Words $B1$ *and* 11 *are words of* B.

Sentential forms

Definition 14.6

For a grammar $G = (N, T, P, S)$, all the words derived from S are called sentential forms.

Let us suppose $\alpha \in (N \cup T)^*$ is such that $S \overset{G*}{\Longrightarrow} \alpha$. α is the sentential form, i.e., α has the potential to become a word in the language of G.

Definition 14.7

If $S \overset{lm*}{\Longrightarrow} \alpha$, we say α is a left-sentential form.

Definition 14.8

If $S \overset{rm*}{\Longrightarrow} \alpha$, we say α is a right-sentential form.

Example: sentential forms

Example 14.10

Consider the following derivations

$$E \Rightarrow E + E \Rightarrow E + B \Rightarrow E \times E + B$$

All above derivations are in sentential forms.

However, if we derive from B we will not be in sentential form.

$$B \Rightarrow B1 \Rightarrow 11$$

The above is not a sentential form.

Example : palindromes

Example 14.11

Consider language of palindromes, i.e., words that are read same in reverse direction. A CFG for the language is

1. $P \to 0P0$
2. $P \to 1P1$

3. $P \to \epsilon$
4. $P \to 1$
5. $P \to 0$

Exercise 14.2

Give a CFG for language $\{ww^R | w \in \Sigma^*\}$

End of Lecture 14