W6 Context-Free Languages (1)

Context-Free Languages

1. Context-Free Grammar (CFG)

- **Finite automata** accept precisely the strings in the language.

Perform a computation to determine whether a specific string is in the language.

- **Regular expressions** describe precisely the strings in the language

Describe the general shape of all strings in the language.

- **Context-free grammar (CFG)** is an entirely different formalism for defining a class

of languages.

Give a procedure for listing off all strings in the language.

Context-Free Grammar

Example

- Start variable S with rules:

  $S \rightarrow AB$

  $A \rightarrow a$

  $A \rightarrow aA$

  $B \rightarrow b$

  $B \rightarrow bB$

variables: S, A, B terminals: a, b

- Following these rules, we can yield ?

Context-Free Grammar

Definition

A context-free grammar is a 4-tuple $G = (V, \Sigma, R, S)$, where

1. V is a finite set, whose elements are called **variables**,
2. Σ is a finite set, whose elements are called **terminals**,
3. $V \cap \Sigma = \emptyset$,
4. S is an element of V ; it is called the **start variable**,
5. R is a finite set, whose elements are called **rules**. Each rule has the form $A \rightarrow w$, where $A \in V$ and $w \in (V \cup \Sigma)^*$.

Context-Free Grammar

Example

Language $L = \{0^k0^k : k \geq 0\}$ has CFG $G = (V, \Sigma, R, S)$,

yeild（捕获量）

Let $G = (V, \Sigma, R, S)$ be a context free grammar with

- $A \in V$
- $u, v, w \in (V \cup \Sigma)^*$,
- $A \rightarrow w$ is a rule of the grammar

The string uwv can be derived in one step from the string uAv, written as

$$uAv \Rightarrow uwv$$

Example: $aaAbb \Rightarrow aaaAbb$

derive（起源）

Let $G = (V, \Sigma, R, S)$ be a context free grammar with

- $u, v \in (V \cup \Sigma)^*$

The string v can be derived from the string u, written as $u \overset{*}{\Rightarrow} v$, if one of the following conditions holds:

1. $u = v$
2. there exist an integer $k \geq 2$ and a sequence $u_1, u_2, \ldots, u_k$ of strings in $(V \cup \Sigma)^*$, such that

(a) $u = u_1$,

(b) $v = u_k$, and $u_1 \Rightarrow u_2 \Rightarrow \ldots \Rightarrow u_k$.

Example: With the rules $A \rightarrow B1 \mid D0C$

$$0AA \overset{*}{\Rightarrow} 0D0CB1$$

2. Language of CFG

Language of CFG

Definition

The language of CFG $G = (V, \Sigma, R, S)$ is

$$L(G) = \{ w \in \Sigma^* \mid S \overset{*}{\Rightarrow} w \}.$$

Such a language is called **context-free**, and satisfies $L(G) \subseteq \Sigma^*$.

Example

CFG $G = (V, \Sigma, R, S)$ with

1. $V = \{S\}$
2. $\Sigma = \{0, 1\}$
3. Rules R: $S \rightarrow 0S \mid \varepsilon$

$L(G) = ?$

Palindrome

Example (Palindrome)

CFG $G = (V, \Sigma, R, S)$ with

1. $V = \{S\}$
2. $\Sigma = \{a, b\}$
3. Rules R: $S \rightarrow aSa \mid bSb \mid a \mid b \mid \varepsilon$

Language of this CFG ?

Simple Arithmetic Expressions

Example (Simple Arithmetic Expressions)

CFG $G = (V, \Sigma, R, S)$ with

1. $V = \{S\}$
2. $\Sigma = \{+, -, \times, /, (,), 0, 1, 2, \ldots, 9\}$
3. Rules R:

   $S \rightarrow S + S \mid S - S \mid S \times S \mid S/S \mid (S) \mid -S \mid 0 \mid 1 \mid \cdot \quad \cdot \quad \cdot \mid 9$

$L(G)$: valid arithmetic expressions over single-digit integers

S derives string $3 \times (5 + 6)$?

Regular Languages are context-free

Theorem

Let Σ be an alphabet and let $L \subseteq \Sigma^*$ be a regular language. Then L is a context-free language (Every regular language is context-free).

Proof

Since L is a regular language, there exists a deterministic finite automaton $M = (Q, \Sigma, \delta, q, F)$ that accepts L. To prove that L is context-free, we have to define a context-free grammar $G = (V, \Sigma, R, S)$, such that $L = L(M) = L(G)$. Thus, G must have the following property:

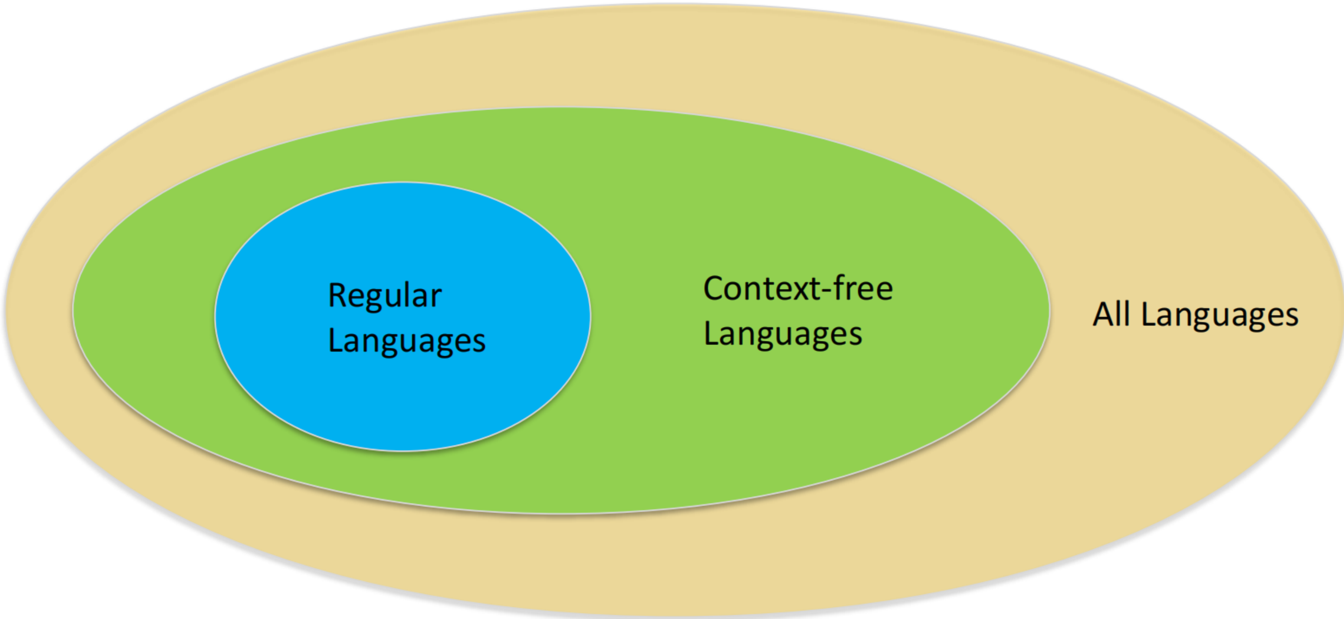
For every string $w \in \Sigma^*$,

$$w \in L(M) \text{ if and only if } w \in L(G),$$

which can be reformulated as

$$M \text{ accepts } w \text{ if and only if } S \overset{*}{\Rightarrow} w.$$

Set $V = \{R_i \mid q_i \in Q\}$ (that is, G has a variable for every state of M). Now, for every transition $\delta(q_i , a) = q_j$ add a rule $R_i \rightarrow aR_j$. For every accepting state $q_i \in F$ add a rule $R_i \rightarrow \varepsilon$. Finally, make the start variable $S = R_0$.

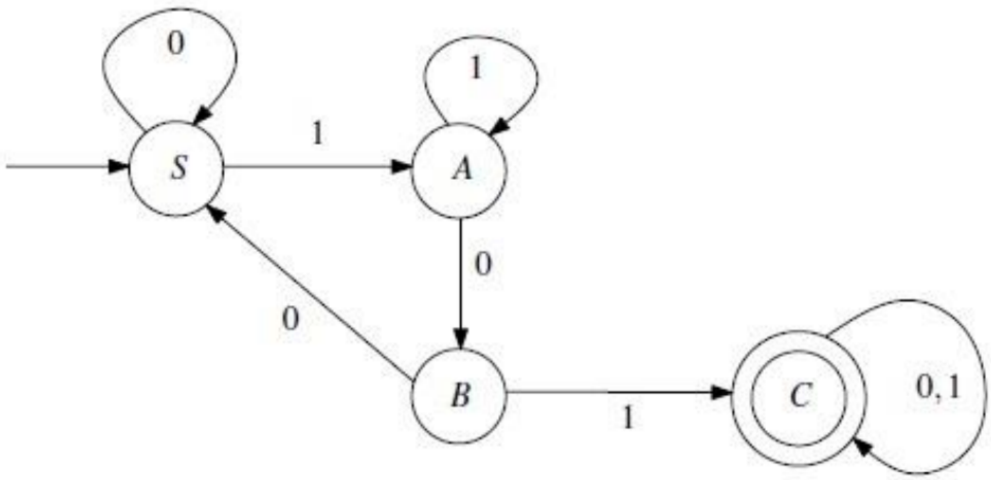


Example

Let L be the language defined as

$$L = \{w \in \{0, 1\}^*: 101 \text{ is a substring of } w\}.$$

The DFA M that accepts L



How can we convert M to a context-free grammar G whose language is L?

3. Chomsky Normal Form (CNF乔姆斯基范式)

Definition

A context-free grammar $G = (V, \Sigma, R, S)$ is said to be in **Chomsky normal form**, if every rule in R has one of the following three forms:

- $A \to BC$, where A, B, and C are elements of V , $B \neq S$, and $C \neq S$.
- $A \to a$, where A is an element of V and a is an element of Σ.
- $S \to \varepsilon$, where S is the start variable.

Why CNF?

Grammars in Chomsky normal form are far easier to analyze.

Example

Rules of CFG in Chomsky normal form with $V = \{S, A, B\}, \Sigma = \{a, b\}$:

G_1: $S \to AB, S \to c, A \to a, B \to b$ (CNF)

G_1: $S \to aA, A \to a, B \to c$ (not CNF)

Rules of CFG in Chomsky normal form with $V = \{S, A, B\}, \Sigma = \{a, b\}$:

G_1: $S \to AB, S \to c, A \to a, B \to b$ (CNF)

G_1: $S \to aA, A \to a, B \to c$ (not CNF)

1. 对于第一个文法 G_1：
   1. $S \to AB$ 符合 $A \to BC$ 形式
   2. $S \to c$ 符合 $A \to a$ 形式
   3. $A \to a$ 符合 $A \to a$ 形式
   4. $B \to b$ 符合 $A \to a$ 形式

因此，第一个文法 G_1 符合乔姆斯基范式（CNF）。

2. 对于第二个文法 G_1：
   1. $S \to aA$ 这是一个非终结符生成一个终结符和另一个非终结符的形式，不符合乔姆斯基范式，因为它不符合 $A \to BC$ 或 $A \to a$ 的形式。
   2. $A \to a$ 符合 $A \to a$ 形式
   3. $B \to c$ 符合 $A \to a$ 形式

因此，第二个文法 G_1 不符合乔姆斯基范式（CNF），因为产生式 $S \to aA$ 不是乔姆斯基范式允许的形式。

Chomsky Normal Form (CNF)

Theorem

Let Σ be an alphabet and let $L \subseteq \Sigma^*$ be a context-free language. There exists a context-free grammar in Chomsky normal form, whose language is L (Every CFL can be described by a CFG in CNF).

CFL $\rightarrow$ CNF

Given CFG $G = (V, \Sigma, R, S)$. Replace, one-by-one, every rule that is not "Chomsky".

- Start variable (not allowed on RHS of rules)
- ε-rules ($A \rightarrow \varepsilon$ not allowed when A isn't start variable)
- all other violating rules ($A \rightarrow B$, $A \rightarrow aBc$, $A \rightarrow BCDE$)

Chomsky Normal Form (CNF)

Theorem

Let Σ be an alphabet and let $L \subseteq \Sigma^*$ be a context-free language. There exists a context-free grammar in Chomsky normal form, whose language is L (Every CFL can be described by a CFG in CNF).

CFL $\rightarrow$ CNF

Given CFG $G = (V, \Sigma, R, S)$. Replace, one-by-one, every rule that is not "Chomsky".

- Start variable (not allowed on RHS of rules)
- ε-rules ($A \rightarrow \varepsilon$ not allowed when A isn't start variable)
- all other violating rules ($A \rightarrow B$, $A \rightarrow aBc$, $A \rightarrow BCDE$)

让Σ是一个字母表，让L⊆Σ□是一个上下文无关的语言。在乔姆斯基范式中存在一个与上下文无关的语法，其语言为L（每个CFL都可以用CNF中的一个CFG来描述）

Converting CFG into CNF

Transformation steps

Step 1. Eliminate the start variable from the right-hand side of the rules.

- New start variable S_0
- New rule $S_0 \to S$

Step 2. Remove **ε-rules** $A \to \varepsilon$, where $A \in V - \{S\}$.

- Before: $B \to xAy$ and $A \to \varepsilon \mid \cdot \cdot \cdot$
- After: $B \to xAy \mid xy$ and $A \to \cdot \cdot \cdot$

When removing $A \to \varepsilon$ rules, insert all new replacements:

- Before: $B \to AbA$ and $A \to \varepsilon \mid \cdot \cdot \cdot$
- After: $B \to AbA \mid bA \mid Ab \mid b$ and $A \to \cdot \cdot \cdot$

Step 3. Remove **unit rules** $A \to B$, where $A \in V$.

- Before: $A \to B$ and $B \to xCy$
- After: $A \to xCy$ and $B \to xCy$

Step 4. Eliminate all rules having more than two symbols on the right-hand side.

- Before: $A \to B_1B_2B_3$
- After: $A \to B_1A_1$, $A_1 \to B_2B_3$

Step 5. Eliminate all rules of the form $A \to ab$, where a and b are not both variables.

- Before: $A \to ab$
- After: $A \to B_1B_2$, $B_1 \to a$, $B_2 \to b$.

Converting CFG into CNF

Example

Given a CFG $G = (V, \Sigma, R, S)$, where $V = \{A, B\}$, $\Sigma = \{0, 1\}$, A is the start variable, and R consists of the rules:

$$A \rightarrow BAB \mid B \mid \varepsilon$$
$$B \rightarrow 00 \mid \varepsilon$$

Convert this G to CNF:

Step 1. Eliminate the start variable from the right-hand side of the rules.

Step 2. Remove **ε-rules**.

(1) Remove $A \rightarrow \varepsilon$: $S \rightarrow A$, $A \rightarrow BAB$

(2) Remove $B \rightarrow \varepsilon$: $A \rightarrow BAB$, $A \rightarrow B$, $A \rightarrow BB$

Step 3. Remove **unit-rules**.

(1) Remove $A \rightarrow A$:

(2) Remove $S \rightarrow A$:

Step 3. Remove **unit-rules**.

(3) Remove $S \rightarrow B$:

(4) Remove $A \rightarrow B$:

Example

Step 4. Eliminate all rules having more than two symbols on the right-hand side.

(1) Remove $S \rightarrow BAB$:

(2) Remove $A \rightarrow BAB$:

Step 5. Eliminate all rules, whose right-hand side contains exactly two symbols, which are not both variables.

(1) Remove $S \rightarrow 00$:

(1) Remove $A \rightarrow 00$:

Step 5. Eliminate all rules, whose right-hand side contains exactly two symbols, which are not both variables.

(3) Remove $S \rightarrow 00$: