Module 1 - Introduction to Languages and Grammars

Lecture Topic

Overview of a Computational Models

Contents

1. Chomsky Hierarchy
2. Basic terminologies
3. Strings and its operations
4. Languages and its operations

Chomsky Hierarchy

	Language	Grammar	Machine	Example
Type 3	Regular languages	Regular grammars • Right-linear grammars • Left-linear grammars	Finite-state automata	a^*
Type 2	Context-free languages	Context-free grammars	Push-down automata	$a^n b^n$
Type 1	Context-sensitive languages	Context-sensitive grammars	Linear-bound automata	$a^n b^n c^n$
Type 0	Recursive languages Recursively enumerable languages	Unrestricted grammars	Turing machines	any computable function

Basic terminologies

Theory of Computation

- **Symbol -** A symbol (often also called a **character**) is the smallest building block, which can be any alphabet, letter, or picture.

- **Alphabet (Σ)** : Finite, nonempty set of symbols which are always ***finite***.
  - Examples:
    - $\sum$ = {0, 1}: binary alphabet
    - $\sum$ = {*a*, *b*, *c*, …, *z*}: the set of all lower case letters
    - The set of all ASCII characters

- **String -** A string is a ***finite*** sequence of symbols from some alphabet $\sum$**.** A string is generally denoted as ***w*** and the length of a string is denoted as **|*w*|**.

  - Examples:
    - 01101 where $\sum$ = {0, 1}
    - *abracadabra* where $\sum$ = {*a*, *b*, *c*, ..., *z*}

Theory of Computation

- **Empty String**: The string with **zero** occurrences of symbols from Σ and is denoted **e** or λ

- **Length of String**: Number of symbols in the string
  - The length of a string w is usually written $|w|$
  - $|1010| = 4$
  - $|\mathcal{E}| = 0$
  - $|uv| = |u| + |v|$

- **Reverse** : w^R
  - If w = abc, w^R = cba

- **Concatenation**: if $\boldsymbol{x}$ and $\mathbf{y}$ are strings, then $\boldsymbol{xy}$ is the string obtained by placing a copy of $\mathbf{y}$ immediately after a copy of $\boldsymbol{x}$
  - $x = a_1a_2 \ldots a_i,\ y = b_1b_2 \ldots b_j$
  - $xy = a_1a_2 \ldots a_ib_1b_2 \ldots b_j$
  - Example: $x = 01101$, $y = 110$, $xy = 01101110$
  - $x\mathcal{E} = \mathcal{E}x = x$

Theory of Computation

Example

- Number of Strings (of length 2)
- that can be generated over the alphabet {a, b}:

  \- \-

  a a

  a b

  b a

  b b

- Length of String |w| = 2
- Number of possible Strings = 4

Therefore

- For alphabet {a, b} with length n, number of strings can be generated **= 2 ^{n}.**
- Let Σ = {a, b} be an alphabet; then aa, ab, bba, baaba, . . . are some examples of strings over Σ.

string operations

Closure Representation

- **L^+**: It is a ***Positive Closure*** that represents a set of all strings except Null or ε-strings.
- **L^***: It is "**Kleene Closure**", that represents the occurrence of certain alphabets for given language alphabets from zero to the infinite number of times. In which ε-string is also included.
- Let $\Sigma = \{a, b\}$ be an alphabet; then aa, ab, bba, baaba, . . . are some examples of strings over Σ.
- The set of all strings over an alphabet Σ is denoted by $\Sigma*$.

$$\mathbf{L^* = \varepsilon L+}$$

string operations

Example:

(a) Regular expression for language accepting all combination of g's over Σ={g}:

$$R = g^*$$

R={ε,g,gg,ggg,gggg,ggggg,...}

(b) Regular Expression for language accepting all combination of g's over Σ={g} : $R = g^+$

R={g, gg, ggg, gggg, ggggg, gggggg,...}

Example:

Input String: "GFG".

Σ^* = { ε,"GFG","GGFG","GGFG","GFGGGGGGGG","GGGGGGGGFFFFFFFFGGGGGGGG",...}

- For example, if Σ = {0, 1}, then $\Sigma *$ = {ε, 0, 1, 00, 01, 10, 11, 000, 001, . . .}.

Languages

Language

A language is a *set of strings*, chosen from some Σ^* or we can say- '**A language is a subset of Σ^* '**.A language that can be formed over ' Σ ' can be **Finite** or **Infinite**. Note : , $|\{\varepsilon\}| = 1$.

- A language is a subset of Σ^*
  - Example of languages:
    - The set of valid Arabic words
    - The set of strings consisting of *n* **0**'s followed by *n* **1**'s
      - $\{\varepsilon$, 01, 0011, 000111, …$\}$
    - The set of strings with equal number of **0**'s and **1**'s
      - $\{\varepsilon$, 01, 10, 0011, 0101, 1010, 1001, 1100, …$\}$
- **Empty language**: $\emptyset = \{ \}$
- The language $\{\varepsilon\}$ consisting of the empty string
- Note: $\emptyset \neq \{\varepsilon\}$

Languages

Language

Since languages are sets, we can apply various well known set operations such as union, intersection, complement, difference on languages.

The notion of concatenation of strings can be extended to languages as follows.

1. The concatenation of a pair of languages L1, L2 is

$$L1.L2 = \{xy \mid x \in L1 \wedge y \in L2\}.$$

Example 2.2.3.

1. If L1 = {0, 1, 01} and L2 = {1, 00}, then L1.L2 = {01, 11, 011, 000, 100, 0100}.

2. For L1 = {b, ba, bab} and L2 = {ε, b, bb, abb}, we have L1.L2 = {b, ba, bb, bab, bbb, babb, baabb, babbb, bababb}.

Strings and Languages (4)

- **Power of an Alphabet**: Σ^k = the set of strings of length k with symbols from S
  - Example: Σ = {0, 1}
  - $S^1 = S = \{0, 1\}$
  - $S^2 = \{00, 01, 10, 11\}$
  - $S^0 = \{e\}$

- **Question**: How many strings are there in S^3?

- The set of all strings over Σ is denoted Σ^*
  - $S^* = S^0 \cup S^1 \cup S^2 \cup S^3 \cup \ldots$
- Also
  - $S^+ = S^1 \cup S^2 \cup S^3 \cup \ldots$
  - $S^* = S^+ \cup \{e\}$
  - $S^+ = S^* - \{e\}$

Strings and Languages (5)

- **Substring**: any string of consecutive characters in some string *w*
  - If *w* = abc
  - e, a, ab, abc are substrings of *w*

- **Prefix** and **suffix**:
  - if w = vu
  - v is a prefix of w
  - u is a suffix of w
  - Example
    - If w = abc
    - a, ab , abc are prefixes of w
    - c, bc, abc are suffixes of w

- Suppose: S is the string banana
  - **Prefix : ban, banana**
  - **Suffix : ana, banana**
  - **Substring : nan, ban, ana, banana**

Basic terminologies

1. Since concatenation of strings is associative, so is the concatenation of languages.

That is, for all languages L1, L2 and L3,

(L1.L2).L3 = L1.(L2.L3).

Hence, (L1L2)L3 may simply be written as L1.L2.L3.

2. The number of strings in L1L2 is always less than or equal to the product of individual numbers, i.e.

$|L1.L2| \leq |L1|.|L2|$.

3. $L1 \subseteq L1.L2$ if and only if $\varepsilon \in L2$.

4. Similarly, $\varepsilon \in L1$ if and only if $L2 \subseteq L1.L2$.

Basic terminologies

1. Kleene star of the language {01} is {ε, 01, 0101, 010101, . . .} = $\{(01)^n / \mid n \geq 0\}$.

2. If L = {0, 10}, then L^* = {ε, 0, 10, 00, 010, 100, 1010, 000, . . .}

Since an arbitrary string in L^n is of the form $x^1 x^2 \cdots x^n$, for xi ∈ L and

$L^* = [\ n\geq 0 \ \ L^n$, one can easily observe that $L* = \{x^1 x^2 \cdots x^n \mid n \geq 0$ and $x_i \in L$, for $1 \leq i \leq n\}$

Thus, a typical string in L∗ is a concatenation of finitely many strings of L.

3. Note that, the Kleene star of the language L = {0, 1} over the alphabet Σ = {0, 1} is

$L * = L^0 \cup L^1 \cup L^2 \cup \cdots$

= {ε} ∪ {0, 1} ∪ {00, 01, 10, 11} ∪ · · ·

= {ε, 0, 1, 00, 01, 10, 11, · · · }

= the set of all strings over Σ.

Basic terminologies

4. The positive closure of a language L is denoted by L^+ is defined as L^+ = Union of L^2 where n>=1

$$\text{Thus, } L^* = L^+ \cup \{\varepsilon\}$$

Exercise :

1. The set of all strings over {a, b, c} that have ac as substring can be written as $\{xacy \mid x, y \in \{a, b, c\}^*\}$.
2. The set of all strings over {a, b} in which every occurrence of b is not before an occurrence of a.
3. The set of all strings over some alphabet Σ with equal number of a 0 s and b 0 s
4. The set of all strings over some alphabet Σ with no consecutive a 0 s.
5. The set of all strings over some alphabet Σ that have an a in the 5th position from the right

Basic terminologies

String Properties:

- P1 Recall that concatenation of languages is associative.
- P2 Since concatenation of strings is not commutative, we have $L1.L2 \neq L2.L1$, in general.
- P3 $L\{\varepsilon\} = \{\varepsilon\}L = L$.
- P4 $L\emptyset = \emptyset L = \emptyset$. $\emptyset L = \emptyset$ as well.
- P5 Distributive Properties:
  1. $(L1 \cup L2)L3 = L1L3 \cup L2L3$.
  2. $L1(L2 \cup L3) = L1L2 \cup L1L3$.
- P6 If $L1 \subseteq L2$ and $L3 \subseteq L4$, then $L1L3 \subseteq L2L4$.
- P7 $\emptyset * = \{\varepsilon\}$.
- P8 $\{\varepsilon\} * = \{\varepsilon\}$.
- P9 If $\varepsilon \in L$, then $L * = L +$.
- P10 $L * L = LL* = L +$.

Strings and Languages (3)

- **Concatenation**: if $\boldsymbol{x}$ and $\boldsymbol{y}$ are strings, then $\boldsymbol{xy}$ is the string obtained by placing a copy of $\boldsymbol{y}$ immediately after a copy of $\boldsymbol{x}$
  - $x = a_1a_2 \ldots a_i,\ y = b_1b_2 \ldots b_j$
  - $xy = a_1a_2 \ldots a_ib_1b_2 \ldots b_j$
  - Example: x = 01101, y = 110, xy = 01101110
  - $xe = ex = x$

Strings and Languages (4)

- **Power of an Alphabet**: Σ^k = the set of strings of length k with symbols from Σ
  - Example: $\Sigma = \{0, 1\}$
  - $\Sigma^1 = \Sigma = \{0, 1\}$
  - $\Sigma^2 = \{00, 01, 10, 11\}$
  - $\Sigma^0 = \{\varepsilon\}$

- **Question**: How many strings are there in Σ^3?

- The set of all strings over Σ is denoted Σ^*
  - $\Sigma^* = \Sigma^0 \cup \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$
- Also
  - $\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \Sigma^3 \cup \ldots$
  - $\Sigma^* = \Sigma^+ \cup \{\varepsilon\}$
  - $\Sigma^+ = \Sigma^* - \{\varepsilon\}$

Strings and Languages (5)

- **Substring**: any string of consecutive characters in some string *w*
  - If *w* = abc
  - e, a, ab, abc are substrings of *w*

- **Prefix** and **suffix**:
  - if w = vu
  - v is a prefix of w
  - u is a suffix of w
  - Example
    - If w = abc
    - a, ab , abc are prefixes of w
    - c, bc, abc are suffixes of w

Strings and Languages (6)

- Suppose: S is the string banana
  - **Prefix : ban, banana**
  - **Suffix : ana, banana**
  - **Substring : nan, ban, ana, banana**

Strings and Languages (7)

- **Language**: set of strings chosen from some alphabet
- A language is a subset of Σ^*
  - Example of languages:
    - The set of valid Arabic words
    - The set of strings consisting of *n* **0**'s followed by *n* **1**'s
      - $\{\varepsilon$, 01, 0011, 000111, …$\}$
    - The set of strings with equal number of **0**'s and **1**'s
      - $\{\varepsilon$, 01, 10, 0011, 0101, 1010, 1001, 1100, …$\}$
- **Empty language**: $\varnothing = \{\ \}$
- The language $\{\varepsilon\}$ consisting of the empty string
- Note: $\varnothing \neq \{\varepsilon\}$

Strings and Languages (8)

- Can concatenate languages
  - $L_1L_2 = \{xy \mid x \in L_1, y \in L_2\}$
  - L^n = L concatenated with itself n times
  - $L^0 = \{\varepsilon\}$; $L^1 = L$

- Star-closure
  - $L^* = L^0 \cup L^1 \cup L^2 \cup \ \ldots$
  - $L^+ = L^* - L^0$

- Languages can be finite or infinite
  - $L = \{a, aba, bba\}$
  - $L = \{a^n \mid n > 0\}$

Strings and Languages (9)

OPERATION	DEFINITION
union of L and M written $L \cup M$	$L \cup M = \{s \mid s \text{ is in } L \text{ or } s \text{ is in } M\}$
concatenation of L and M written LM	$LM = \{st \mid s \text{ is in } L \text{ and } t \text{ is in } M\}$
Kleene closure of L written L*	$L^* = \cup\, L^i$, i=0,.., ∞ L* denotes “zero or more concatenations of “ L
positive closure of L written L^+	$L^+ = \cup\, L^i$, i=1,.., ∞ L^+ denotes “one or more concatenations of “ L $L^+ = LL^*$

Strings and Languages (10)

$L = \{A, B, \ldots, Z, a, b, \ldots z\}$ $\quad D = \{1, 2, \ldots, 9\}$

$L \cup D$ = the set of letters and digits

LD = all strings consisting of a letter followed by a digit

L^2 = the set of all two-letter strings

$L^4 = L^2\ L^2$ = the set of all four-letter strings

L^* = { All possible strings of L plus ε }, $L^+ = L^* - \varepsilon$

D^+ = set of strings of one or more digits

$L\,(L \cup D)$ = set of all strings consisting of a letter followed by a a letter or a digit

$L\,(L \cup D)^*$ = set of all strings consisting of letters and digits beginning with a letter

Strings and Languages (11)

- The language **L** consists of strings over **{a,b}** in which **each** string begins with an **a** should have an **even** length

  - aa, ab $\in$ L
  - aaaa,aaab,aaba,aabb,abaa,abab,abba,abbb $\in$ L

  - baa $\notin$ L
  - a $\notin$ L

Strings and Languages (12)

- The language **L** consists of strings over **{a,b}** in which each occurring of **b** is **immediately preceded** by an **a**

  - $\varepsilon \in L$
  - $a \in L$
  - $abaab \in L$

  - $bb \notin L$
  - $bab \notin L$
  - $abb \notin L$

Strings and Languages (13)

- Let $X = \{a,b,c\}$ and $Y = \{abb, ba\}$. Then
  - $XY = \{aabb, babb, cabb, aba, bba, cba\}$
  - $X^0 = \{\varepsilon\}$
  - $X^1 = X = \{a,b,c\}$
  - $X^2 = XX = \{aa,ab,ac,ba,bb,bc,ca,cb,cc\}$
  - $X^3 = XXX =$

    $\{aaa,aab,aac,aba,abb,abc,aca,acb,acc,baa,bab,bac,bba,bbb,bbc,bca,bcb,bcc,caa,cab,$ $cac,cba,cbb,cbc,cca,ccb,ccc\}$

Strings and Languages (14)

- The language $L = \{a,b\}^*\{bb\}\{a,b\}^* = \Sigma^*\{bb\}\Sigma^*$
  - consists of the strings over $\{a,b\}$ that contain the substring **bb**
  - $bb \in L$
  - $abb \in L$
  - $bbb \in L$
  - $aabb \in L$
  - $bbaaa \in L$
  - $bbabba \in L$
  - $abab \notin L$
  - $bab \notin L$
  - $b \notin L$

Strings and Languages (15)

- Let **L** be the language that consists of all strings that begin with **aa** or end with **bb**
  - $L_1 = \{aa\}\{a,b\}^*$
  - $L_2 = \{a,b\}^*\{bb\}$
  - $L = L_1 \cup L_2 = \{aa\}\{a,b\}^* \cup \{a,b\}^*\{bb\}$
  - $bb \in L$
  - $abb \in L$
  - $bbb \in L$
  - $aabb \in L$
  - $bbaaa \notin L$
  - $bbabba \notin L$
  - $abab \notin L$
  - $bab \notin L$
  - $ba \notin L$

Strings and Languages (16)

- Let $\mathbf{L_1}$ = {bb} and $\mathbf{L_2}$ = {ε, bb, bbbb} over **b**

- The languages L_1^* and L_2^* both contain precisely the strings consisting of an even number of **b**'s.

- ε , with length zero, is an element of L_1^* and L_2^*

Strings and Languages (17)

- What is the language of all even-length strings over **{a,b}**
  - L = $\{aa, bb, ab, ba\}^*$ = $(aa|bb|ab|ba)^*$

- What is the language of all odd-length strings over **{a,b}**
  - L = $\{a,b\}^*$ - $\{aa, bb, ab, ba\}^*$ or
  - L = $\{a,b\}\{aa, bb, ab, ba\}^*$ or
  - L = $\{aa, bb, ab, ba\}^* \{a,b\}$