Tianjin International Engineering Institute

Formal Languages and Automata

Lesson 2: Formal Languages

Marc Gaetano

Edition 2018

Alphabets and strings

- A common way to talk about words, number, pairs of words, etc. is by representing them as strings
- To define strings, we start with an alphabet

An alphabet is a finite set of symbols

- Examples

$\Sigma_1 = \{a, b, c, d, \ldots, z\}$: the set of letters in English

$\Sigma_2 = \{0, 1, \ldots, 9\}$: the set of (base 10) digits

$\Sigma_3 = \{a, b, \ldots, z, \#\}$: the set of letters plus the special symbol #

$\Sigma_4 = \{\, (, \,) \,\}$: the set of open and closed brackets

Strings

A string over alphabet Σ is a finite sequence of symbols in Σ

- The empty string will be denoted by ε
- Examples

abfbz is a string over $\Sigma_1 = \{a, b, c, d, \ldots, z\}$

9021 is a string over $\Sigma_2 = \{0, 1, \ldots, 9\}$

ab#bc is a string over $\Sigma_3 = \{a, b, \ldots, z, \#\}$

))()(() is a string over $\Sigma_4 = \{\ (,\ )\ \}$

Languages

A language is a set of strings over an alphabetC

- Languages can be used to describe problems with "yes/no" answers, for example:

$L_1 =$ The set of all strings over Σ_1 that contain the substring "fool"

$L_2 =$ The set of all strings over Σ_2 that are divisible by 7 $= \{7, 14, 21, \ldots\}$

$L_3 =$ The set of all strings of the form $s\#s$ where s is any string over $\{a, b, \ldots, z\}$

$L_4 =$ The set of all strings over Σ_4 where every (can be matched with a subsequent)

Set membership

Computation is translated to set membership

Example computation problem:

Is number x prime?

Equivalent set membership problem:

$$x \in PRIMES = \{2, 3, 5, 7, 11, 13, 17, \ldots\}?$$

String Operations

$w = a_1 a_2 \cdots a_n$ $\qquad$ $abba$

$v = b_1 b_2 \cdots b_m$ $\qquad$ $bbbaaa$

Concatenation

$wv = a_1 a_2 \cdots a_n b_1 b_2 \cdots b_m$ $\qquad$ $abbabbbaaa$

String Operations

$$w = a_1 a_2 \cdots a_n$$

$$ababaaabbb$$

Reverse

$$w^R = a_n \cdots a_2 a_1$$

$$bbbaaababa$$

String Operations

$$w = a_1 a_2 \cdots a_n$$

$$abba$$

$$aa$$

$$a$$

Length

$$|w| = n$$

$$|abba| = 4$$

$$|aa| = 2$$

$$|a| = 1$$

Length and Concatenation

$$|uv| = |u| + |v|$$

Example:

$$u = aab \;, \quad |u| = 3$$

$$v = abaab \;, \quad |v| = 5$$

$$|uv| = |aababaab| = 8$$

$$|uv| = |u| + |v| = 3 + 5 = 8$$

Empty String

A string with no letters is denoted: ε

Acts as a neutral element

Observations:

$$|\varepsilon| = 0$$

$$\varepsilon w = w\varepsilon = w$$

$$\varepsilon abba = abba\varepsilon = ab\varepsilon ba = abba$$

Substring

Substring of string:

a subsequence of consecutive characters

String	Substring
$\underline{ab}bab$	ab
$\underline{abba}b$	$abba$
$ab\underline{b}ab$	b
$a\underline{bbab}$	$bbab$

Prefix and Suffix

string $abbab$

Prefixes	Suffixes
ε	$abbab$
a	$bbab$
ab	bab
abb	ab
$abba$	b
$abbab$	ε

$$w = \underbrace{u}_{\text{prefix}}\underbrace{v}_{\text{suffix}}$$

Exponent Operation

$$w^n = \underbrace{ww \cdots w}_{n}$$

Example: $(abba)^2 = abbaabba$

Definition: $w^0 = \varepsilon$ $\quad (abba)^0 = \varepsilon$

The * Operation

Σ^* is the set of all possible strings from alphabet Σ

$$\Sigma = \{a, b\}$$

$$\Sigma^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, aab, \ldots\}$$

The + Operation

Σ^{+} is the set of all possible strings from alphabet Σ except ε

$$\Sigma = \{a, b\}$$

$$\Sigma^{*} = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, aab, \ldots\}$$

$$\Sigma^{+} = \Sigma^{*} - \varepsilon$$

$$\Sigma^{+} = \{a, b, aa, ab, ba, bb, aaa, aab, \ldots\}$$

Language

Any language over alphabet Σ is a subset of Σ^*

Example:

$$\Sigma = \{a, b\}$$

$$\Sigma^* = \{\varepsilon, a, b, aa, ab, ba, bb, aaa, \dots\}$$

Languages:

$$\{\}$$

$$\{\varepsilon\}$$

$$\{a, aa, aab\}$$

$$\{\varepsilon, abba, baba, aa, ab, aaaaaa\}$$

Language Example

$\Sigma = \{a, b\}$ $\quad L = \{a^n b^n : n \geq 0\}$

An infinite language

$$\left.\begin{array}{l} \varepsilon \\ ab \\ aabb \\ aaaaabbbbb \end{array}\right\} \in L$$

$bbabb \notin L$

$abb \notin L$

Two special languages

Empty language

$\{ \ \}$ or $\varnothing$

Language with empty string

$\{\varepsilon\}$

Size of a language (number of elements):

$$|\{\}| = 0$$

$$|\{\varepsilon\}| = 1$$

$$|\{a, aa, ab\}| = 3$$

$$|\{\varepsilon, aa, bb, abba, baba\}| = 5$$

Two special languages

Sets $\emptyset = \{\ \} \neq \{\varepsilon\}$

Set size $\left|\{\ \}\right| = \left|\emptyset\right| = 0$

Set size $\left|\{\varepsilon\}\right| = 1$

String length $\left|\varepsilon\right| = 0$

Operations on Languages

The usual set operations:

$\{a, ab, aaaa\} \cup \{bb, ab\} = \{a, ab, bb, aaaa\}$ union

$\{a, ab, aaaa\} \cap \{bb, ab\} = \{ab\}$ intersecti on

$\{a, ab, aaaa\} - \{bb, ab\} = \{a, aaaa\}$ difference

Complement: $\overline{L} = \Sigma^* - L$

$$\overline{\{a, ba\}} = \{\varepsilon, b, aa, ab, bb, aaa, \ldots\}$$

Reverse

Definition:

$$L^R = \{ w^R : w \in L \}$$

Examples:

$$\{ ab, aab, baba \}^R = \{ ba, baa, abab \}$$

$$L = \{ a^n b^n : n \geq 0 \}$$

$$L^R = \{ b^n a^n : n \geq 0 \}$$

Concatenation

Definition:

$$L_1 L_2 = \{ xy : x \in L_1, y \in L_2 \}$$

Example:

$$\{a, ab, ba\}\{b, aa\}$$

$$\{ab, aaa, abb, abaa, bab, baaa\}$$

Power operation

Definition:

$$L^n = \underbrace{LL \cdots L}_{n}$$

Exemple:

$$\{a, b\}^3 = \{a, b\}\{a, b\}\{a, b\} =$$

$$\{aaa, aab, aba, abb, baa, bab, bba, bbb\}$$

Special case:

$$L^0 = \{\varepsilon\}$$

Star-Closure (Kleene *)

Definition:

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

Example:

$$\{a, bb\}^* = \left\{ \begin{array}{l} \varepsilon, \\ a, bb, \\ aa, abb, bba, bbbb, \\ aaa, aabb, abba, abbbb, \ldots \end{array} \right\} \quad \begin{array}{l} L^0 \\ L^1 \\ L^2 \\ L^3 \end{array}$$

Positive closure (+)

Definition:

$$L^{+} = L^{1} \cup L^{2} \cup L^{3} \cup \cdots$$

Example:

$$\{a, bb\}^{+} = \left\{ \begin{array}{l} a, bb, \\ \\ aa, abb, bba, bbbb, \\ \\ aaa, aabb, abba, abbbb, \ldots \end{array} \right\} \begin{array}{l} L^{1} \\ \\ L^{2} \\ \\ L^{3} \end{array}$$

Note that: $L^{*} = L^{0} \cup L^{+}$