形式语言与自动机理论

课程简介与基础知识

丁效

xding@ir.hit.edu.cn

计算学部

哈尔滨工业大学

2023年2月

任课教师介绍

任课教师：丁效

教授/博士生导师

研究方向：人工智能、自然语言处理
文本推理、知识计算

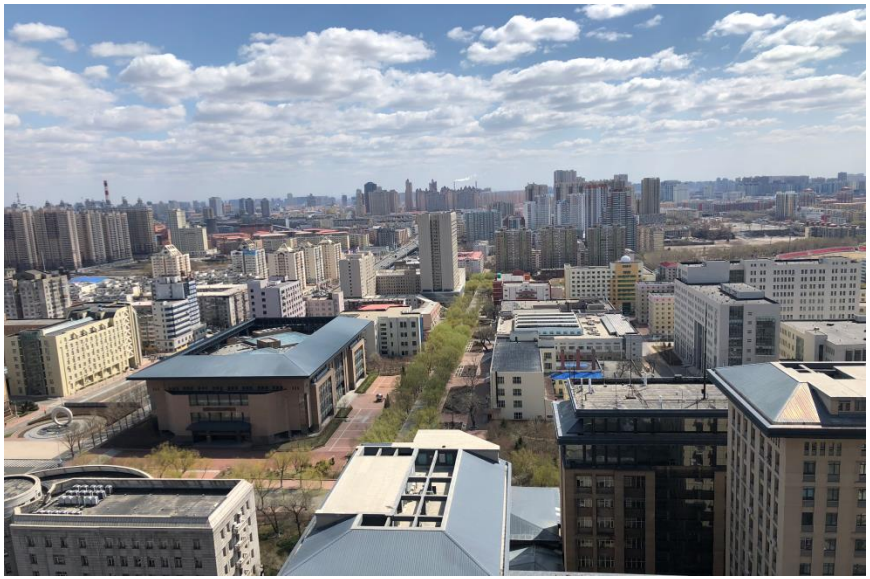
单　　位：计算学部计算机科学与技术学院
社会计算与信息检索研究中心**(SCIR)**

办公地点：科学园科创大厦K1304室

联系方式：**13845078754　xding@ir.hit.edu.cn**

Formal Language and Automata

- 语言
- 自动机
- 计算

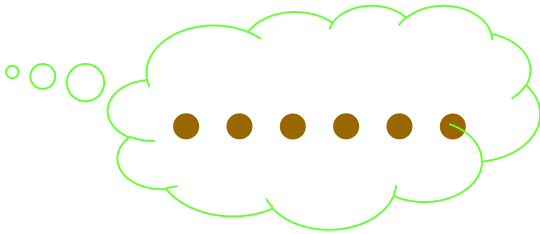


语言

What is a language ?

This is a sentence.

This is also a sentence.



So we have

{ sentence 1, sentence 2, sentence 3, ... }

the set of sentences ⇔ Language

形式语言

形式语言: 经数学定义的语言

		自然语言		形式语言	
		English	中文	化学分子式	C语言
语言	字符	A, a, B, b,…	天，地，...	A-Z, a-z, 0-9 ...	A-Z, a-z, 0-9 ...
	单词	Apple	苹果	H_2O	char
	句子	How are you?	早上好!	$2H_2+O_2=2H_2O$	char a = 10;
	语法	Grammar	语法规则	精确定义的规则	

字母表

字母表： 符号 (或字符) 的非空有穷集合

$\Sigma_1 = \{ 0, 1 \}$

$\Sigma_2 = \{a, b, \ldots, z \}$

$\Sigma_3 = \{x | x是一个汉字\}$

symbols $\Rightarrow$ strings $\Rightarrow$ language

字符串

字符串：由某字母表中符号组成的有穷序列

(1 + 2) * (13 - 7)

To stay at home and save lives.

不聚集，戴口罩，勤洗手。

0, 1, 00, 01, 10, 000, 001, 1010, 00111100

空串：记为ε，有0个字符的串

字母表Σ可以是任意的，但是都要求$\varepsilon \notin \Sigma$

字符串的长度

字符串的长度:字符串中符号所占位置的个数，记为$|\cdot|$

若字母表为 Σ, 可递归定义为:

$$|w| = \begin{cases} 0 & w = \varepsilon \\ |x| + 1 & w = xa \end{cases}$$

其中 a ∈ Σ, w 和 x 是 Σ 中字符组成的字符串

字符串的长度

字符串的长度:字符串中符号所占位置的个数，记为$|\cdot|$

若字母表为 Σ, 可递归定义为:

$$|w| = \begin{cases} 0 & w = \varepsilon \\ |x| + 1 & w = xa \end{cases}$$

其中 a ∈ Σ, w 和 x 是 Σ 中字符组成的字符串

- 符号使用的一般约定:
  - 字母表: Σ, Γ, …
  - 字符: a, b, c, …
  - 字符串: …, w, x, y, z
  - 集合: A, B, C, …

字符串的连接

字符串 x 和 y 的**连接:** 将首尾相接得到新串的运算, 记为 x·y 或 xy

同样, 可递归定义为:

$$x \cdot y = \begin{cases} x & y = \varepsilon \\ (x \cdot z)a & y = za \end{cases}$$

其中 a ∈ Σ, x, y, z都是字符串

对任何字符串 x, 有 ε· x = x ·ε = x.

连接运算的符号 “·” 一般省略

字符串的幂

字符串 x 的n次**幂**(n≥0)，递归定义为：

$$x^n = \begin{cases} \varepsilon & n = 0 \\ x^{n-1}x & n > 0 \end{cases}$$

其中，x是字符串

字符串 x 的n次**幂**可以理解为将字符串x重复n次

集合的连接

集合 A 和 B 的**连接:** 记为A·B或 AB

$A \cdot B = \{w \mid w = x \cdot y , x \in A \text{ 且 } y \in B\}$

集合的幂

集合 A 的n次**幂**(n≥0)，递归定义为：

$$A^n = \begin{cases} \{\varepsilon\} & n = 0 \\ A^{n-1}A & n \geq 1 \end{cases}$$

集合的幂

集合 A 的n次**幂**(n≥0)，递归定义为：

$$A^n = \begin{cases} \{\varepsilon\} & n = 0 \\ A^{n-1}A & n \geq 1 \end{cases}$$

那么, 若 Σ 为字母表, 则 Σ^n 为 Σ 上长度为 n 的字符串集合。如果 Σ = {0,1}, 有

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

$\Sigma^1 = \{0, 1\}$

$\Sigma^2 = \{00, 01, 10, 11\}$

$\Sigma^3 = \{000, 001, 010, 011, 100, 101, 110, 111\}$

克林闭包(Kleene Closure)

$$\Sigma^* = \bigcup_{i=0}^{\infty} \Sigma^i$$

正闭包(Positive Closure)

$$\Sigma^+ = \bigcup_{i=1}^{\infty} \Sigma^i$$

显然,

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

语言

定义： 若 Σ 为字母表且 $\forall L \subseteq \Sigma^*$, 则 L 称为字母表 Σ 上的语言

- 自然语言, 程序设计语言等
- $\{0^n1^n \mid n \geq 0\}$
- The set of strings of 0's and 1's with an equal number of each:

  $\{\varepsilon, 01, 10, 0011, 0101, 1100, \ldots\}$

关于语言：唯一重要的约束就是所有字母表都是有穷的

问题

典型问题

判断给定的字符串 w 是否属于某个具体的语言 L，

$w \in L$?

- 任何所谓问题，都可以转为语言成员性的问题
- 语言和问题其实是相同的

规则/语法

语法：生成字符串的规则

英文语法举例:

<sentence> → <noun-phrase><predicate>

<noun-phrase> → <article><noun>

<predicate> → <verb>

<article> → a|an|the

<noun> → wolf|sheep

<verb> → love|eat

I have a dream.

Example 1

字母表：{0,1}

L={w|w consists of 0's and 1's,

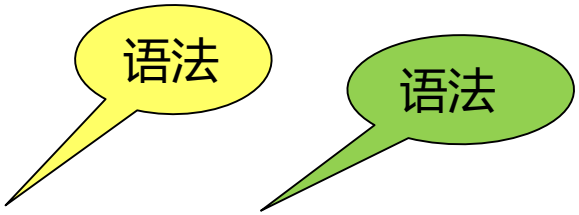
and end with 0 }

语法

L={ 0,00,10,000,010,100,110,0000,……}

11111100∈L, 1∉L, 0001∉L, 20∉L

Example 2



$L=\{ 0^n1^n \mid n \geq 0 \}$

字母表 = {0,1}

$L=\{\varepsilon, 01, 0011, 000111, 00001111, 0000011111, \ldots\ldots \}$

空字符串

Example 3

L={ w| w is a sentence in English }

Everyone loves his/her motherland. √

Sheep eat grass . √

Grass eat sheep . ?

♣ 形式语言关注的是字符串的构成方式，而不关注字符串的语义或内涵

字符串操作

$w = a_1 a_2 \ldots\ldots a_m$ abc

$v = b_1 b_2 \ldots\ldots b_n$ 123456

- Concatenation

$wv = a_1 a_2 \ldots\ldots a_m b_1 b_2 \ldots\ldots b_n$ abc123456

$\neq$

$vw = b_1 b_2 \ldots\ldots b_n a_1 a_2 \ldots\ldots a_m$ 123456abc

- Reverse

$w^R = a_m a_{m-1} \ldots\ldots a_1$ cba

语言操作

- Usual set operations

$$L_1 \cup L_2 = \{ w \mid w \in L_1 \text{ or } w \in L_2 \}$$

$$L_1 \cap L_2 = \{ w \mid w \in L_1 \text{ and } w \in L_2 \}$$

$$L_1 - L_2 = \{ w \mid w \in L_1 \text{ and } w \notin L_2 \}$$

- Reverse

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

- Concatenation

$$L_1 L_2 = \{ wv \mid w \in L_1 \text{ and } v \in L_2 \}$$

Example 4

$L = \{ ab, abc, abcd \} \Rightarrow L^R = \{ ba, cba, dcba \}$

$L = \{ a^n b^n \mid n \geq 1 \} \Rightarrow L^R = \{ b^n a^n \mid n \geq 1 \}$

$L = \{ a^n b^n \mid n \geq 1 \}$, $K = \{ 0^n 1^n \mid n \geq 1 \}$

$LK = \{ a^n b^n 0^n 1^n \mid n \geq 1 \}$ ×

$LK = \{ a^n b^n 0^m 1^m \mid n \geq 1 , m \geq 1 \}$

$L^2 = ?$

* / Star Operation on Languages

$\Sigma = \{ 0, 1 \}$

$\Sigma^* = \Sigma^0 \cup \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \ldots\ldots \cup \Sigma^n \cup \ldots\ldots$

$\Sigma^0 = \{ \varepsilon \}$, $\Sigma^n = \underbrace{\Sigma\,\Sigma \ldots\ldots \Sigma}_{n}$

$\{0,1\}^* = \{\varepsilon\} \cup \{0,1\} \cup \{0,1\}^2 \cup \ldots\ldots \cup \{0,1\}^n \cup \ldots\ldots$

$= \{ \varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 111, \ldots\ldots \}$

Empty string / language

Denote ε as empty string

$|\varepsilon| = 0, \quad w\varepsilon = \varepsilon w = w$

Denote ϕ as empty language

$\phi = \{\}, \quad \phi L = L\phi = \phi$

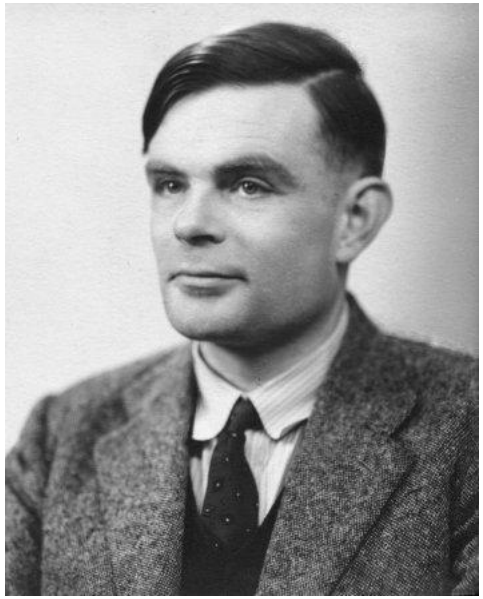
Denote $\Sigma^+ = \Sigma \cup \Sigma^2 \cup \Sigma^3 \cup \ldots\ldots \cup \Sigma^n \cup \ldots\ldots$

$= \Sigma^* - \Sigma^0$

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

自动机

Alan MarthisonTuring



- 研究抽象机器及其所能解决问题的理论

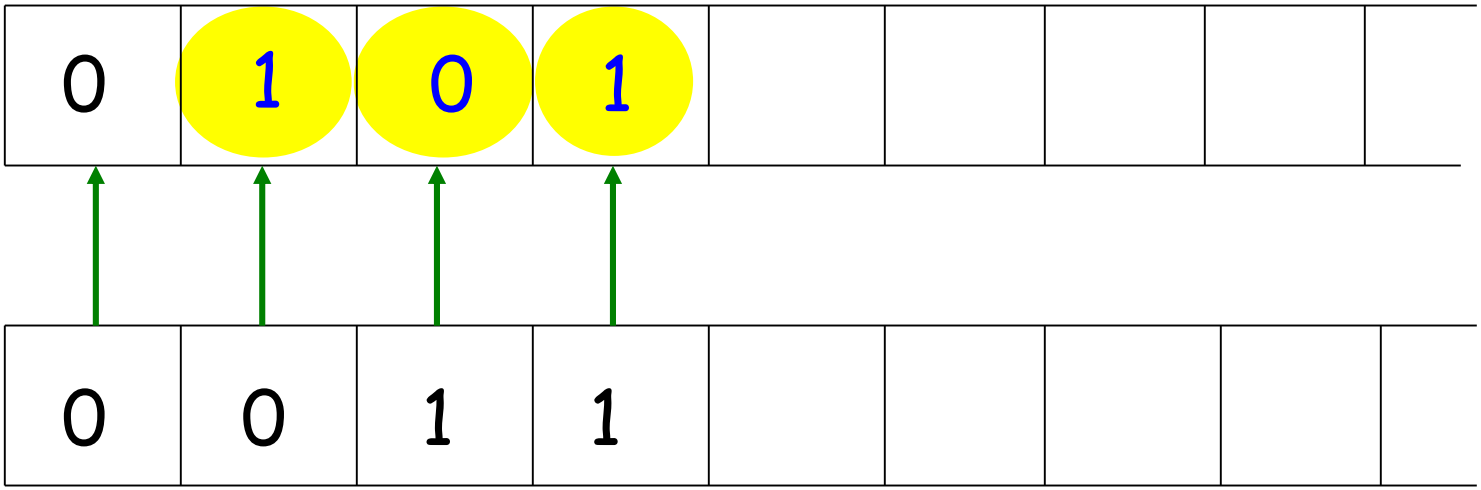
- Turing Machine

自动机

- Finite Automata
  - Deterministic Finite Automata
  - Non-deterministic Finite Automata
- Push Down Automata
- Turing Mashine

计算

$$\begin{array}{r} 2 \\ +\ 3 \\ \hline 5 \end{array} \qquad \begin{array}{r} 0010 \\ +\ 0011 \\ \hline 0101 \end{array}$$



计算

In our life

Coded

In computer

Process

0011101100001011 1100100111100
0011101100001011 11001001111000
001110110000010 11001001111

Compose

Automaton

Language

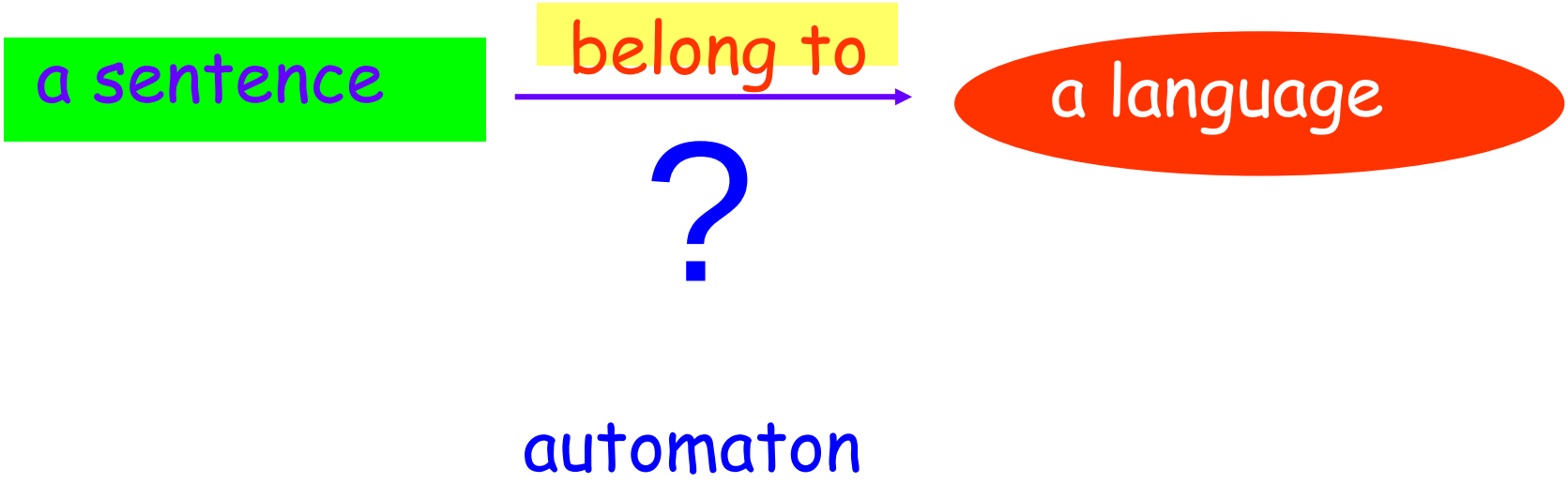
(formal description)

(set of strings)

计算

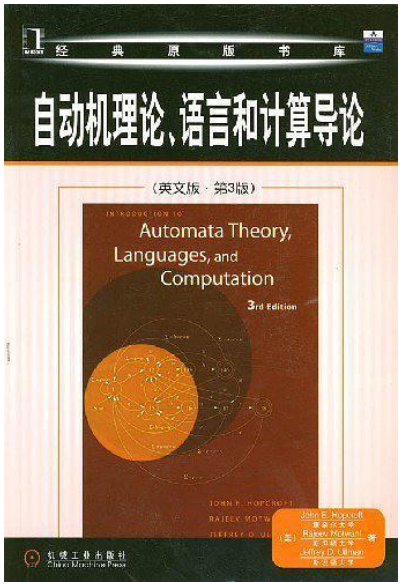
- Computable Problems
  - write a program to solve
- Intractable Problems
  - find someway to work around

Undecidable Problem



Text book

1. Introduction to Automata Theory , Languages , and Computation (Third Edition)



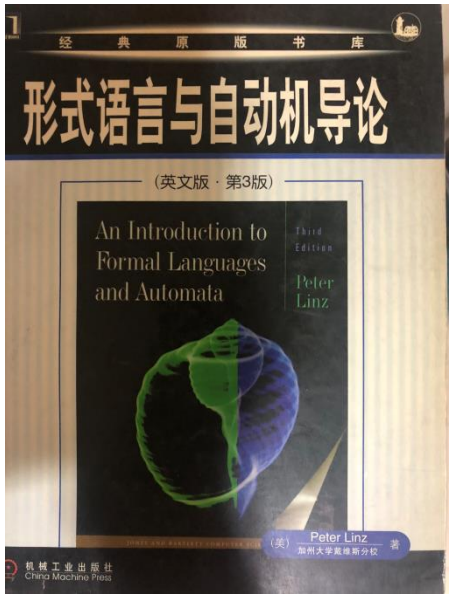
—— John E. Hopcroft

Rajeev Motwani

Jeffrey D. Ullman

Text book

2. An Introduction to Formal Languages and Automata (Third Edition)



—— Peter Linz

Goal

1. Understanding "theoretical" concepts

------ method of formal description

2. Get a sense of how to reason formally

3. Improving reading ability in English

Homework

- All exercises listed on qq-group
- Write on A4 papers
- Submit - nonditermined
- Discussions – maybe

Honor and Collaboration

- Collaboration is strongly encouraged
- Solutions must be written independently
- Responsible for Understanding and explaining

Examination

- Only final exam
- Closed exam

Nothing allowed except one pen

Grading Policy

- Homework : 20% //including Class Performance
- Final exam : 80%

Information

- Tutor ：丁效
- Office：科创大厦K1304
- E-mail： xding@ir.hit.edu.cn
- 课程群： 768304271（qq）

- MOOC：

  https://www.icourse163.org/learn/HIT-1206319802

- 画状态图：http://madebyevan.com/fsm/

Good good study
day day up !