

抽吸引理

pumping lemma

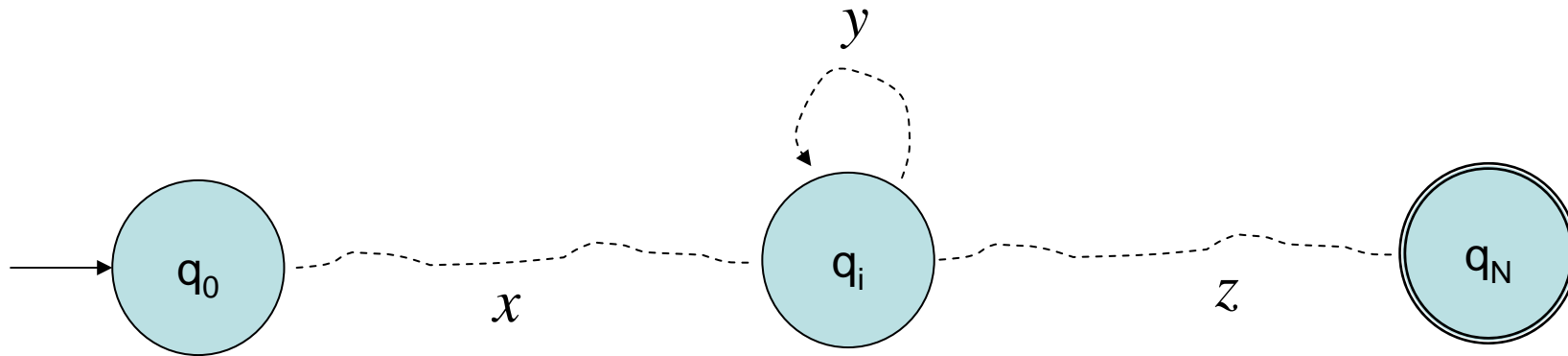
对于给定的有限状态自动机，状态的数目是固定的。因此：

**特征1：** 识别符号串所需要的“记忆容量”不能跟输入字符串长度有比例关系。也就是说，待识别的字符串长度增加，有限状态自动机的“记忆容量”无法相应增大。

**特征2：** 如果一个正则语言具有任意长的符号串（比有限状态自动机的状态数还多），那么该语言所对应的有限状态自动机中一定存在回路（loop）。

《自然语言处理综论》（第13章）中译本p305。

# pumping lemma



$xz$

$xyz$

$xyyz$

$xyyy\dots z$

$xy^n z$

设 $L$ 是正则语言，那么必定存在符号串 $x$ ,  $y$ ,  $z$ , 使得对于 $n \geq 0$ , 有 $y \neq \epsilon$  并且 $xy^n z \in L$

# 证明一种语言不是正则语言

- 如果语言 $L$ 中不能用适当的方法对符号串进行抽吸，那么 $L$ 就不是正则语言。

# 实例1

- 证明： $a^n b^n$  不是正则语言。

$a^n b^n$ 构成的符号串 $s$ ，只可能有三种办法来分割：

1.  $y$ 只有若干 $a$ 组成，因此 $x$ 也是由若干 $a$ 组成， $z$ 包含全部的 $b$ 。如果 $x$ 跟 $z$ 中 $ab$ 个数相等，那么 $xyz$ 中， $a$ 的个数就比 $b$ 的个数多。
2.  $y$ 只由若干 $b$ 组成，这样， $b$ 的个数就比 $a$ 多。
3.  $y$ 由若干 $a$ 和若干 $b$ 组成，这意味着 $x$ 只由 $a$ 组成， $z$ 只由 $b$ 组成，这时候，必然有一些 $b$ 在 $a$ 之前。因为 $y$ 可以重复出现，假定 $y=aaabbb$ ，就会得到 $aaabbbbaaabbb$ ， $aaabbbbaaabbbbaaabbb$ 这样的串。

可见，无论用上面的哪一种方法， $s$ 都不可能分割为 $x$ ， $y$ ， $z$ 三部分，使得 $y$ 能够被抽吸。因此， $s$ 不是正则语言。

## 实例2

- 语言  $L = \{xx^R, x \in a, b^*\}$   $x^R$ 表示 $x$ 的逆向字符串（镜像）。求证： $L$ 不是正则语言。

证明：

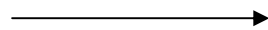
- (1) 令  $RL = aa^*bbaa^*$ ，则  $RL$  是正则语言  
(  $bb$  前后 $a$ 的个数不必相等)
- (2) 令  $L' = RL \cap L = a^n b^2 a^n$  ( $bb$  前后 $a$ 的个数必相等)
- (3) 用抽吸引理可证  $L'$  不是正则语言（思路同例1）
- (4) 所以， $L$  不是正则语言（正则语言的交运算是封闭的）

## 实例2

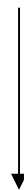
- if  $S_1$ , then  $S_2$
- Either  $S_3$ , or  $S_4$
- The man who said  $S_5$  is arriving today

**If** *either* the man who said  $S_5$  is arriving today *or* the man who said  $S_5$  is arriving tomorrow, **then** the man who said  $S_6$  is arriving the day after ...

if --> a  
then --> a  
either --> b  
or --> b  
other -->  $\epsilon$



a b b a



# 实例3

- The cat likes tuna fish.
- The cat the dog chased likes tuna fish.
- The cat the dog the rat bit chased likes tuna fish.
- The cat the dog the rat the elephant admired bit chased likes tuna fish.

(the noun)<sup>n</sup> (vt)<sup>n-1</sup> likes tuna fish       $n \geq 1$

x = the noun = [the cat, the dog, the rat, the elephant, the tiger, ...]

y = vt = [chased, bit, admired, ate, befriended, ...]

$L = x^n y^{n-1}$

因为  $x^m y^m$  不是正则语言，

是正则语言吗？

因此  $L = x x^m y^m$  ( $m = n-1$ ) 也不是正则语言。