有限状态技术和形态分析

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

什么是形态学(Morphology)？

- 形态学研究屈折语中词的构成规则。
  - 英语、德语等是屈折语(inflectional language)
  - 汉语是孤立语(或分析语)(isolating/analytic language)
  - 日语是黏着语(agglutinative language)
- 词通常由语素(morpheme)组成。
- 语素是语言中最小的意义单位(minimal meaning-bearing unit)。
  - 英语: fox, cats(cat+s), unbelievably(un+believe+able+ly)
  - 德语: gesagt(ge+sagen+t)

英语中的语素

- 总的来说，语素可以分成两大类：
  1. 词干(stem)：提供词的主要意义
     - fox, cats(cat+s), unbelievably(un+believe+able+ly)
  2. 词缀(affix)：提供词的各种附加意义(修改词干义或改变词的语法功能)
     1) 前缀(prefix)：出现在词干的前面
        - rewrites(re+write+s), unbelievable(un+believe+able)
     2) 后缀(suffix)：出现在词干的后面
        - unbelievably(un+believe+able+ly)

英语中的语素

- 语素如何构成词?
  1. 屈折变化(inflection):词干+词缀形成的词通常与原词干同属一类，常用来使词具备数、时态等功能
     - cat+s　walk+ed　walk+ing
  2. 派生(derivation):词干+词缀形成的词通常与原词干不属一类，词义通常与原词干有联系(有时难以预料)
     - computerize+ation (verb→noun)

英语中的屈折变化

- 名词
  1) 单复数
     - cats thrushes boys butterflies (规则变化)
     - mouse/mice, goose/geese, ox/oxen (不规则)
  2) 名词所有格
     - children’s zebra’s　zebras’
- 形容词、副词
  1) 比较级 cleverer　better
  2) 最高级 cleverest best

英语中的屈折变化

- 动词
  1) 一般现在时，单数第三人称 walks eats
  2) 现在分词、动名词 walking eating
  3) 过去式 walked ate
  4) 过去分词 walked eaten
- 规则变化的动词有四种形态

stem / -s form / -ing participle / past form or –ed participle

walk / walks / walking / walked

- 不规则变化的动词有五种形态

stem / -s form / -ing participle / past form / –ed participle

eat / eats / eating / ate / eaten

英语中的屈折变化

- 不规则变化的词数量有限，但多是常用词
- 大部分词的变化属规则变化
  - 规则变化的拼写规则
    - 单个辅音字母结尾，重复该字母再加 ing 或 ed
      - beg begging begged
      - picnic picnicking picnicked
    - 不发音的e结尾，删掉e再加ing 或 ed
      - merge merging merged
    - 以s、z、sh、ch等结尾，加es
      - fax faxes
    - 辅音字母加y结尾，y改写成i加es
      - spy spies

英语中的派生词

- 英语中屈折变化较简单(同德语、俄语等相比)
- 英语中派生现象较为复杂，仅看几个例子
  - 动词、形容词的名词化(nominalization)：
    - computerize (V) → computerization
    - appoint (V) → appointee
    - kill (V) → killer
    - fuzzy (A) → fuzziness
  - 从名词、动词派生出形容词
    - computation (N) → computational
    - adjust (V) → adjustable
    - clue (N) → clueless
- 派生规则规律性不如屈折变化规则、不能随意派生

什么是形态分析？

- 形态分析研究如何利用计算机把屈折语中的词分解成语素。 (morphological parsing)

  foxes→fox+N+PL

  stem + morphological feature

- 形态分析器(morphological parser)
- stemmer/lemmatizer

为什么要进行形态分析

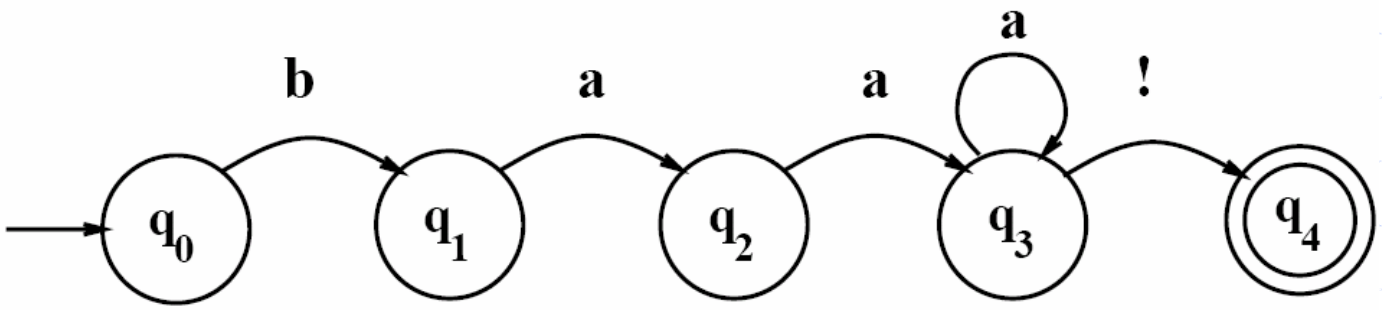
- 为什么不把所有词的形态收进词表？
  - 即使可以，也不效率
  - 有规律，没有必要
  - 对有些形态非常发达的语言(如:土耳其语)，不可能
- 形态分析是应用的要求，如在信息检索中
  - 查询fox当然也希望查到包含foxes的页面
  - 查询relevant 也能想到relevance等
- 形态分析也是深层英语分析的基础

形态分析的基本技术

- 有限状态技术(Finite state technique)
- 回顾一下我们学过的有限状态技术
  - 有限状态自动机(FSA)
  - 有限状态自动机的确定性(DFSA/NFSA)
  - 有限状态自动机和语言
  - 有限状态自动机和正规文法(等价)
  - 有限状态自动机和正则表达式(等价)
  - 有限状态自动机作为正则语言的识别装置和生成装置。

有限状态自动机(Finite State Automaton)

- 下面的有限状态自动机定义语言 baa! baaa! baaaa! …
- 有限状态自动机的构成
  - 一组状态。（例:q_0、q_1、…、q_4）
  - 一组字母(字母表)。（例:b、a、!）
  - 一个开始状态。（例:q_0）
  - 一个或多个终止状态。（例:q_4）
  - 若干状态转换关系。



有限状态自动机的形式定义

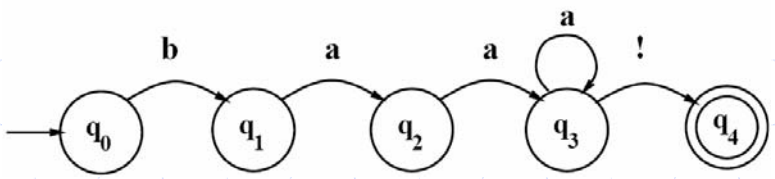
- 一个有限状态自动机M是一个五元组(Q, Σ, q_0, F, δ).
  - 有限个状态组成的状态集: Q
  - 有限字母组成的字母表: Σ
  - 一个开始状态 q_0
  - 终止状态的集合 $F \subseteq Q$
  - 状态转移函数 $\delta(q,i)$: $Q \times \Sigma \rightarrow Q$
- 状态转换图和状态转移矩阵

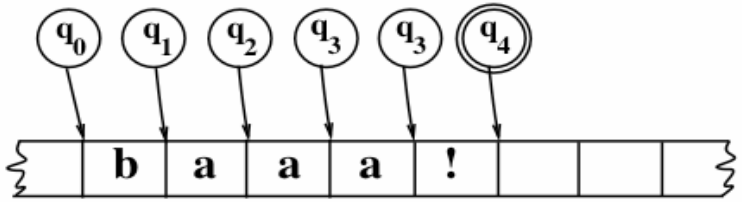
	Input		
State	b	a	!
0	1	∅	∅
1	∅	2	∅
2	∅	3	∅
3	∅	3	4
4:	∅	∅	∅

有限状态自动机作为识别装置

- 可以用来判定一个符号串是否是有限状态自动机所定义的语言中的句子。
- 识别过程
  - 始于开始状态
  - 读入输入指针指针处的输入字母
  - 查询状态转移关系
  - 进入新的状态，移动输入指针位置
  - 是否读完待识别符号串中所有字母？
- 例: 识别baaa!

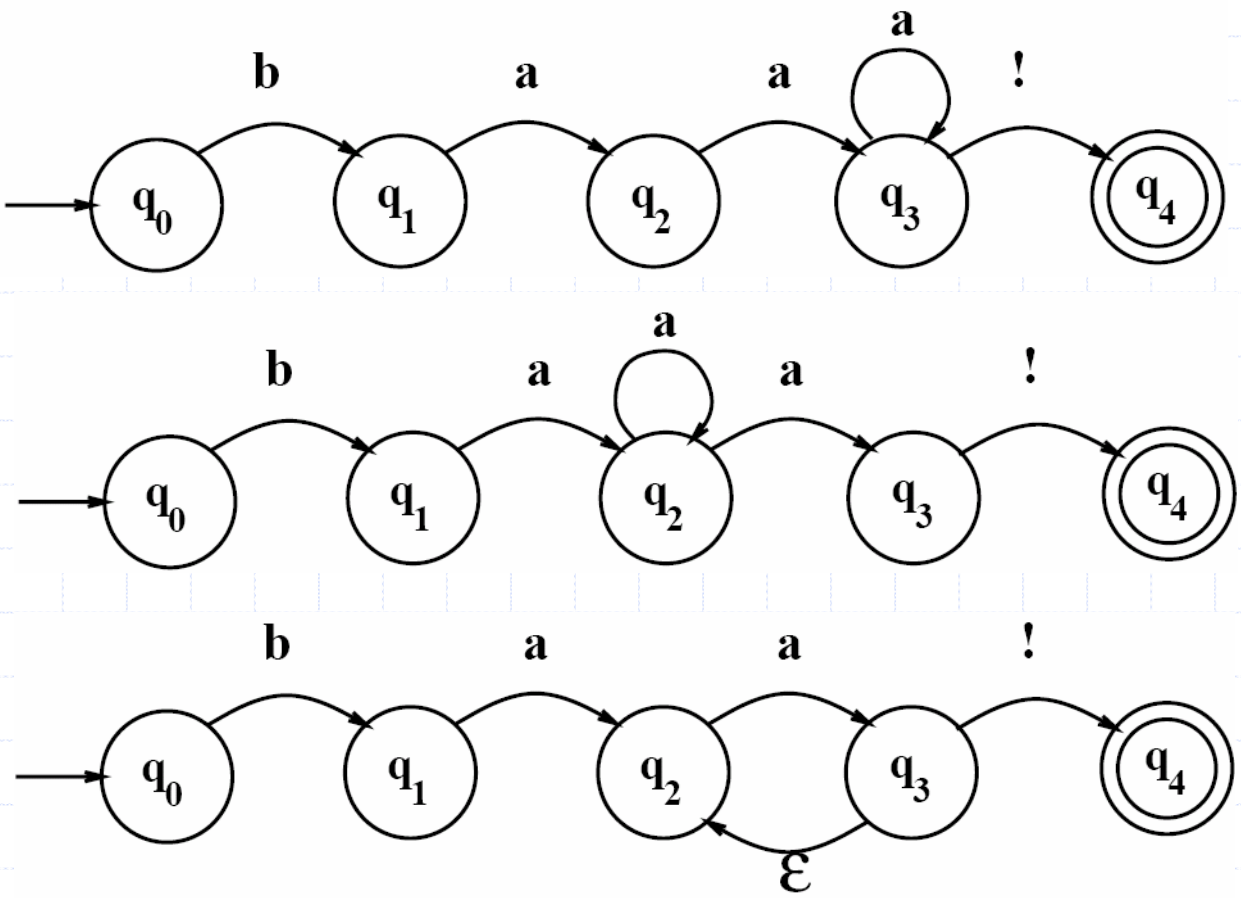
	Input		
State	b	a	!
0	1	Ø	Ø
1	Ø	2	Ø
2	Ø	3	Ø
3	Ø	3	4
4:	Ø	Ø	Ø





非确定的有限状态自动机

- 下面的有限状态自动机识别同样的语言



DFSA和NFSA

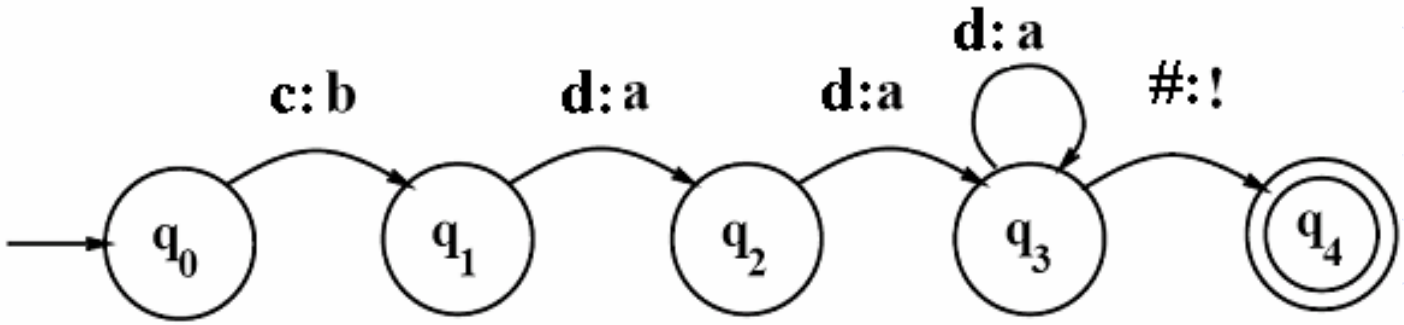
- NFSA的不确定性
  - 同一状态、同一输入字母可转移到多个状态
  - ε弧
- NFSA可以转换成DFSA，所以NFSA不比DFSA能力强。
- NFSA可先转换成DFSA后再进行句子的识别和生成
- 也可以直接用NFSA进行句子的识别和生成，但要处理因此带来的非确定性问题。
- 有时候，使用NFSA更自然。
- 有时候，从NFSA得到的DFSA太复杂(状态多)。

非确定性的处理策略

- 在NFSA中，由于在某个状态，读入同一个输入字母时，可能存在多种选择，因此关键的问题时如何进行选择，如果做了错误的选择，应保证还可以重新进行选择。
- 三种处理策略
  - 引入回溯机制
  - 引入展望符号(look ahead)
  - 引入并行机制

有限状态转换机(FST)

- 如果把FSA中弧上的字母换成两个字母，一个称为输出字母、一个称为输入字母，这样得到FSA就是一个有限状态转换机(Finite State Transducer)。



- 有限状态转换机建立起两个字符串间的联系。

有限状态转换机(FST)

- 一个有限状态转换机M是一个五元组(Q, Σ, q_0, F, δ).
  - 有限个状态组成的状态集: Q
  - 有限个复杂字母组成的字母表: Σ，其中每个复杂字母由一个输出字母和一个输入字母组成，形如 o:i
  - 一个开始状态 q_0
  - 终止状态的集合 $F \subseteq Q$
  - 状态转移函数 $\delta(q,o:i)$: $Q \times \Sigma \rightarrow Q$

有限状态转换机(FST)

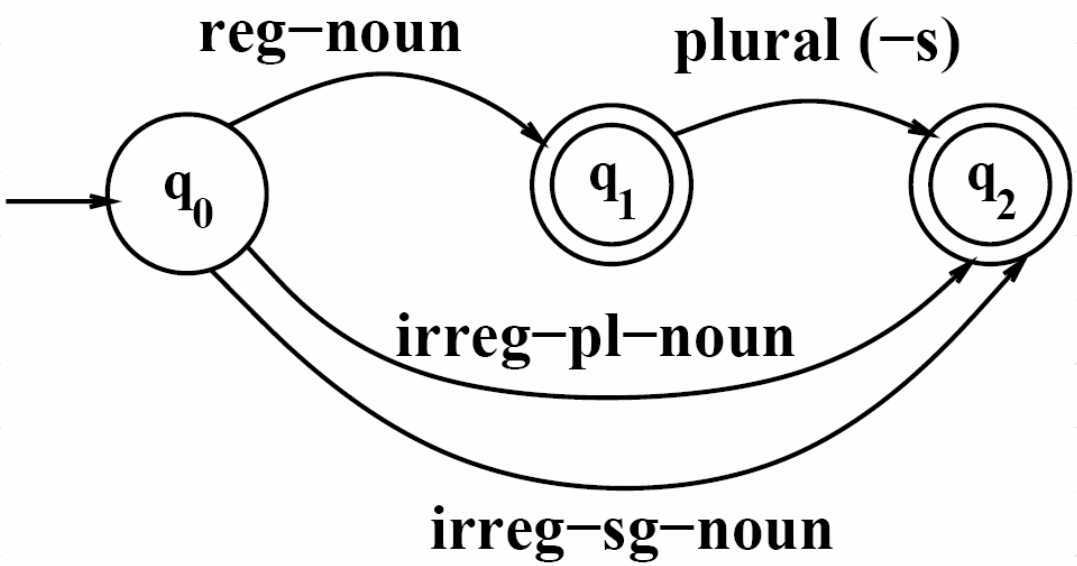
- FST有两条输入带子
- FST作为识别装置(recognizer)
  - 给定一对字符串，FST拒绝或接受
- FST作为生成装置(generator)
  - 生成一对字符串
- FST作为翻译装置(translator)
  - 给定一个字符串，生成另一个字符串

构建形态分析器所需要的资源

1. 词典(lexicon)
   - 词干和词缀
   - 词干和词缀的基本信息(如:词干的类别)
2. 形态知识(morphotactics)
   - 语素间的顺序关系
   - 那一类语素可以和那一类语素组合
     (例如：名词后面可以加一个复数语素)
3. 正字规则(orthographic rule or spelling rule)
   - 两个语素组合时应进行怎样的变化
     (如:把y改写为i加es)

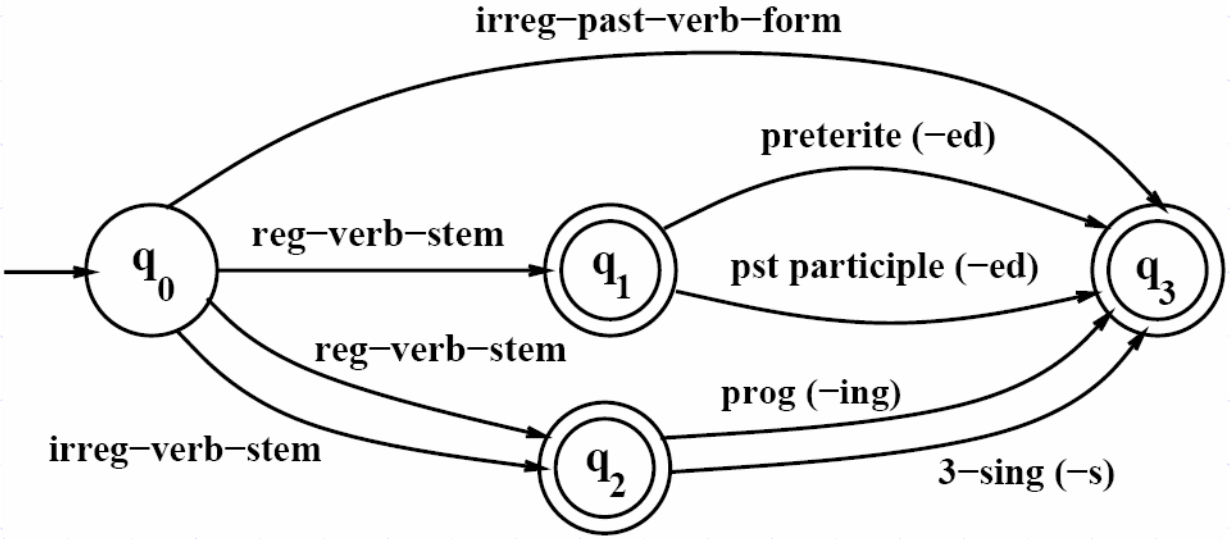
形态知识的表示

- 形态知识(morphotactics)可以通过FSA来表示
- 英语名词屈折变化模型
  - reg-noun
    - fox cat
  - irreg-sg-noun
    - mouse goose
  - irreg-pl-noun
    - mice geese



形态知识的表示

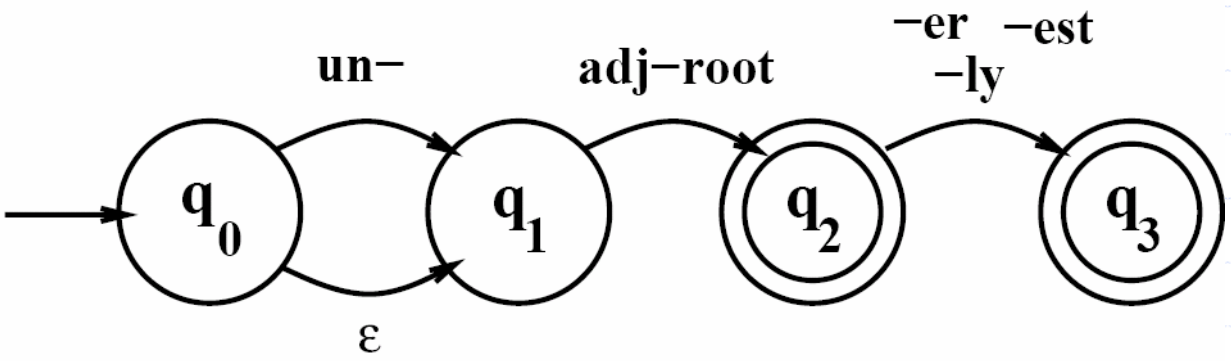
- 英语动词屈折变化模型



reg-verb-stem	walk fry talk impeach
irreg-verb-stem	cut speak sing
irreg-past-verb	caught ate eaten

形态知识的表示

- 形容词屈折和派生变化模型



big bigger biggest

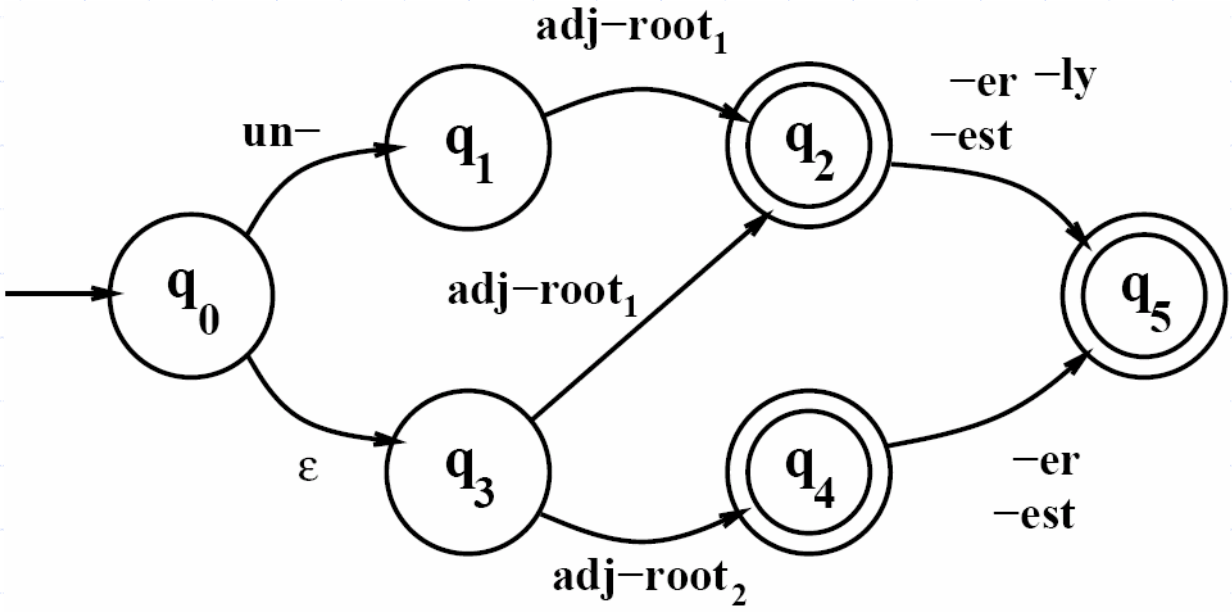
cool cooler coolly

clear clearer clearest clearly unclear unclearly

unbig redly realest （×）

形态知识的表示

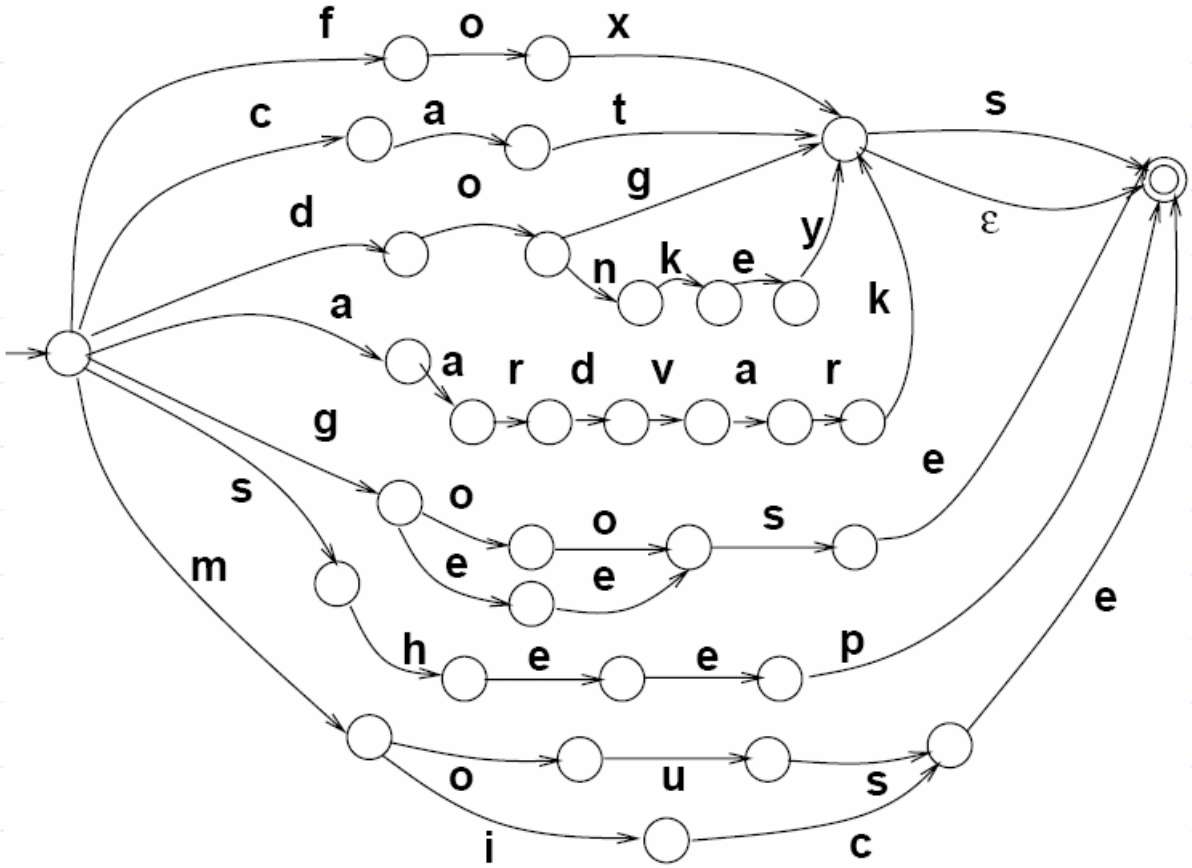
- 对 adj-root 进行分类
  - adj-root$_1$ 可以有前缀 un- 以及后缀 –ly (clear happy)
  - adj-root$_2$ 不可以有前缀 un- 以及后缀 –ly (big red)



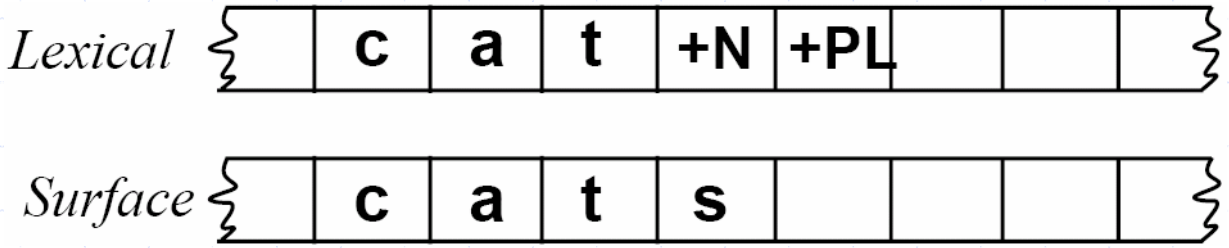
- 派生变化处理起来很复杂，不再介绍

形态知识的表示

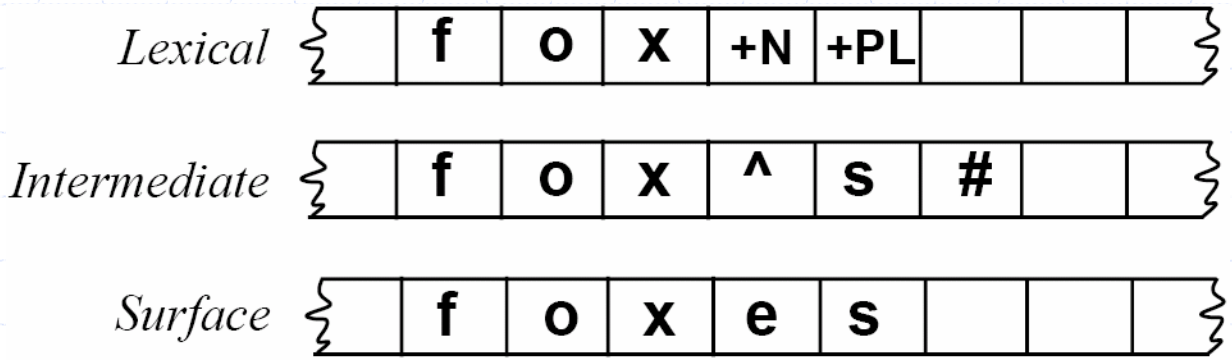
- 用词典中词条取代reg-noun、irreg-sg-noun、irreg-pl-noun等(识别aardvarks)



形态分析



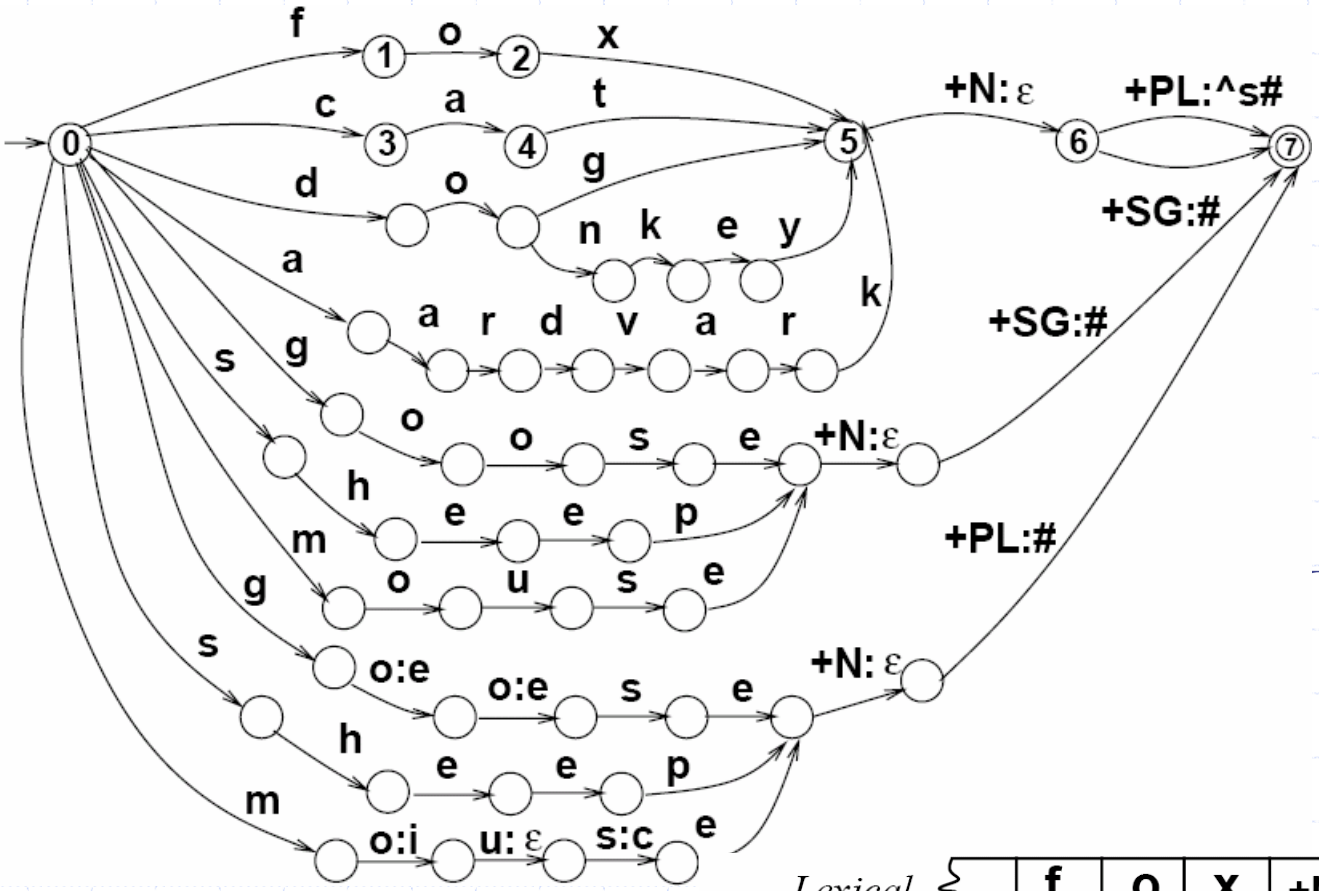
- 二级形态变化



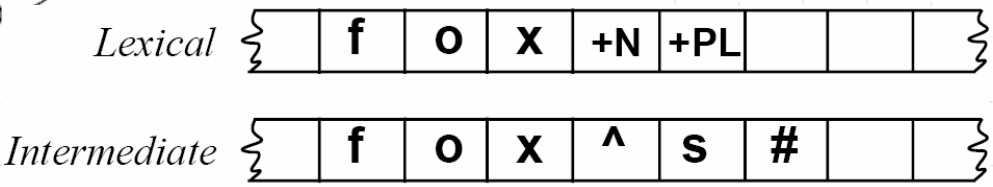
^ morpheme boundary

word boundary

形态分析和FST



一级FST



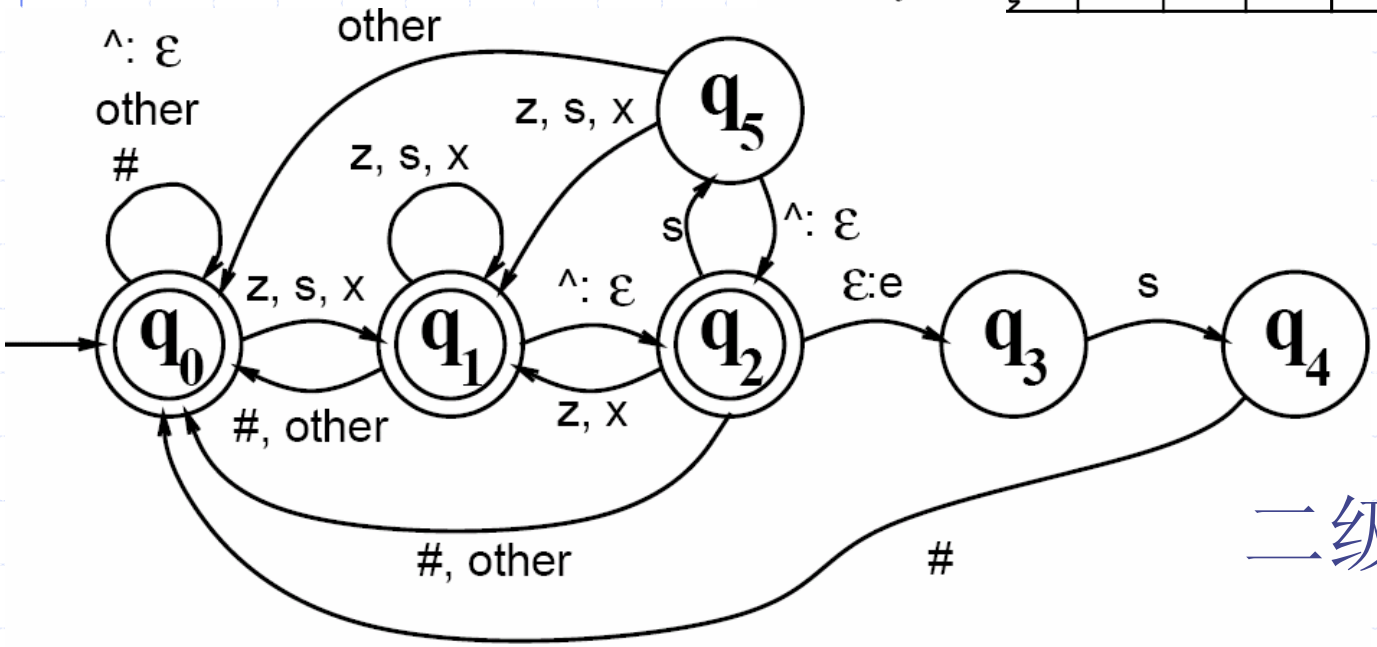
正字规则

- 正字规则(orthographic rule)
  - consonant doubling
    - 1-letter consonant doubled before –ing/-ed
    - e.g. beg/begging
  - E deletion
    - silent e droped before –ing/-ed
    - e.g. make/making
  - E insertion
    - e added after –s –z –x –ch –sh before -s
    - e.g. watch watches
  - Y replacement
    - y changes to –ie before –s, -i before –ed
    - e.g. try tries
  - K insertion
    - verbs ending with vowel + -c add –k before –ing/-ed
    - e.g. panic panicked

E-insertion 和 FST

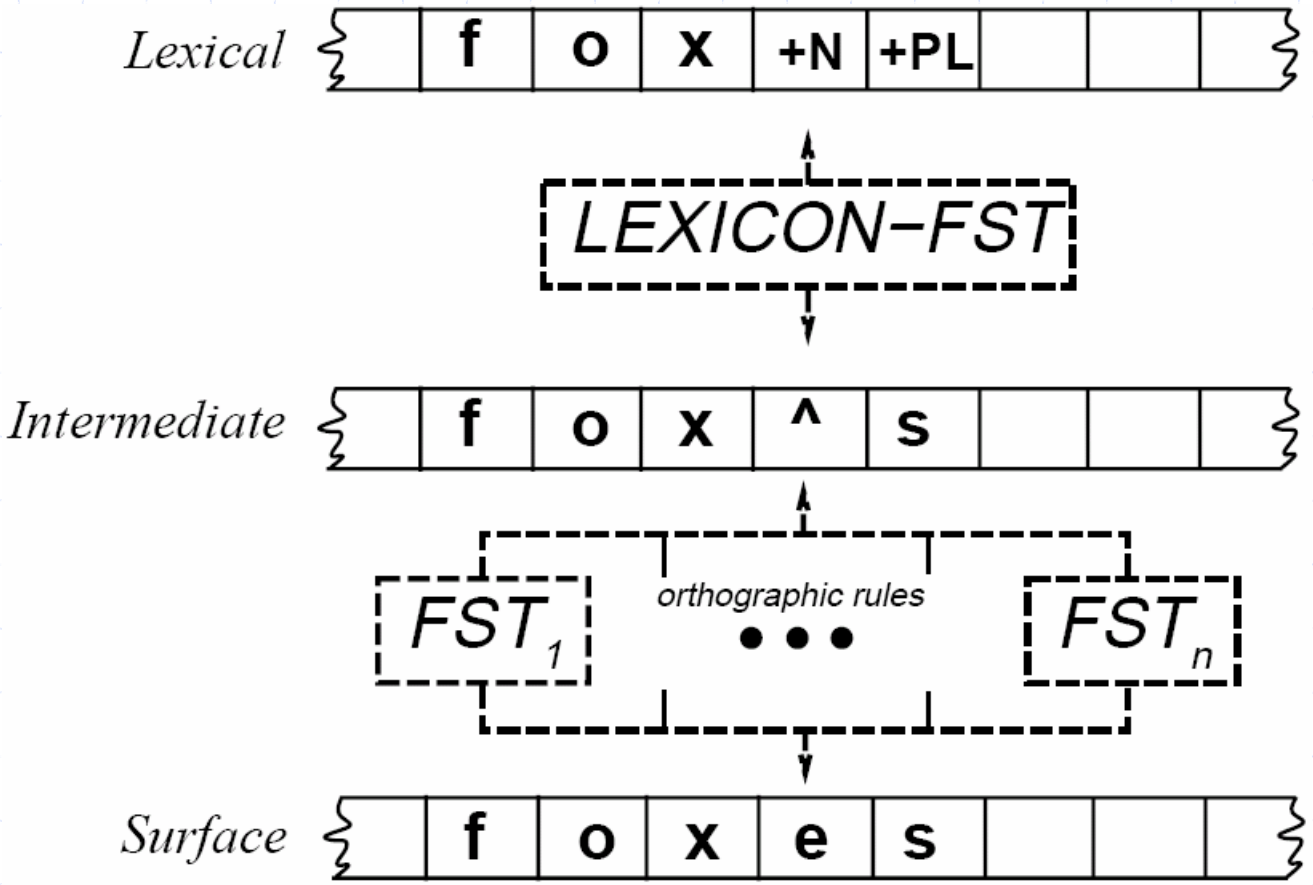
$$\varepsilon \rightarrow e / \left\{ \begin{matrix} x \\ s \\ z \end{matrix} \right\} \wedge \_\_\_ \; s\# \qquad a \rightarrow b / c \_\_\_ d$$

Intermediate	f	o	x	^	s	#		
Surface	f	o	x	e	s			



二级FST

二级形态处理


Lexical
f o x +N +PL
LEXICON−FST
Intermediate
f o x ^ s
FST1
orthographic rules
FSTn
Surface
f o x e s

形态分析

- 算法需要考虑非确定性问题。
- PCKIMMO
  http://www.sil.org/pckimmo/v2/pc-kimmo_v2.html
- XEROX Language Tool
  http://www.xrce.xerox.com/competencies/content-analysis/demos/english