词性标注与语言模型

李辰

cli@xjtu.edu.cn

2024年9月

2/3

课程提纲

一、词性标注

二、语言模型

词性标注

概念

- Part-of-speech (POS) tagging
- 为分词结果中的每个单词标注一个正确的词性。即确定每个词是名词、动词、形容词或其他词性的过程。
- 是高级自然语言处理任务的基础。

举例

人民网/nz 1月1日/t 讯/ng 据/p 《/w [纽约/nsf 时报/n]/nz 》/w 报道/v ，/w 美国/nsf 华尔街/nsf 股市/n 在/p 2013年/t 的/ude1 最后/f 一天/mq 继续/v 上涨/vn ，/w 和/cc [全球/n 股市/n]/nz 一样/uyy ，/w 都/d 以/p [最高/a 纪录/n]/nz 或/c 接近/v [最高/a 纪录/n]/nz 结束/v 本/rz 年/qt 的/ude1 交易/vn 。/w

Mr./NNP Vinken/NNP is/VBZ chairman/NN of/IN Elsevier/NNP N.V./NNP ,/, the/DT Dutch/NNP publishing/VBG group/NN

汉语词性对照表

代码	名称	说明	举例
a	形容词	取英语形容词adjective的第1个字母	最/d 大/a 的/u
ad	副形词	直接作状语的形容词.形容词代码a和副词代码d并在一起	一定/d 能够/v 顺利/ad 实现/v 。/w
ag	形语素	形容词性语素。形容词代码为a，语素代码ｇ前面置以a	喜/v 煞/ag 人/n
an	名形词	具有名词功能的形容词。形容词代码a和名词代码n并在一起	人民/n 的/u 根本/a 利益/n 和/c 国家/n 的/u 安稳/an 。/w
b	区别词	取汉字“别”的声母	副/b 书记/n 王/nr 思齐/nr
c	连词	取英语连词conjunction的第1个字母	全军/n 和/c 武警/n 先进/a 典型/n 代表/n
d	副词	取adverb的第2个字母，因其第1个字母已用于形容词	两侧/f 台柱/n 上/ 分别/d 雄踞/v 着/u
dg	副语素	副词性语素。副词代码为d，语素代码ｇ前面置以d	用/v 不/d 甚/dg 流利/a 的/u 中文/nz 主持/v 节目/n 。/w
e	叹词	取英语叹词exclamation的第1个字母	嗬/e ！/w
f	方位词	取汉字“方” 的声母	从/p 一/m 大/a 堆/q 档案/n 中/f 发现/v 了/u
g	语素	绝大多数语素都能作为合成词的“词根”，取汉字“根”的声母	例如dg 或ag
h	前接成分	取英语head的第1个字母	目前/t 各种/r 非/h 合作制/n 的/u 农产品/n
i	成语	取英语成语idiom的第1个字母	提高/v 农民/n 讨价还价/i 的/u 能力/n 。/w
j	简称略语	取汉字“简”的声母	民主/ad 选举/v 村委会/j 的/u 工作/vn
k	后接成分		权责/n 明确/a 的/u 逐级/d 授权/v 制/k
l	习用语	习用语尚未成为成语，有点“临时性”，取“临”的声母	是/v 建立/v 社会主义/n 市场经济/n 体制/n 的/u 重要/a 组成部分/l 。/w
m	数词	取英语numeral的第3个字母，n，u已有他用	科学技术/n 是/v 第一/m 生产力/n
n	名词	取英语名词noun的第1个字母	希望/v 双方/n 在/p 市政/n 规划/vn
ng	名语素	名词性语素。名词代码为n，语素代码ｇ前面置以n	就此/d 分析/v 时/Ng 认为/v
nr	人名	名词代码n和“人(ren)”的声母并在一起	建设部/nt 部长/n 侯/nr 捷/nr
ns	地名	名词代码n和处所词代码s并在一起	北京/ns 经济/n 运行/vn 态势/n 喜人/a
nt	机构团体	“团”的声母为t，名词代码n和t并在一起	[冶金/n 工业部/n 洛阳/ns 耐火材料/l 研究院/n]nt
nx	字母专名		ＡＴＭ/nx 交换机/n
nz	其他专名	“专”的声母的第1个字母为z，名词代码n和z并在一起	德士古/nz 公司/n
o	拟声词	取英语拟声词onomatopoeia的第1个字母	汩汩/o 地/u 流/v 出来/v
p	介词	取英语介词prepositional的第1个字母	往/p 基层/n 跑/v 。/w
q	量词	取英语quantity的第1个字母	不止/v 一/m 次/q 地/u 听到/v ，/w
r	代词	取英语代词pronoun的第2个字母,因p已用于介词	有些/r 部门/n
s	处所词	取英语space的第1个字母	移居/v 海外/s 。/w
t	时间词	取英语time的第1个字母	当前/t 经济/n 社会/n 情况/n
tg	时语素	时间词性语素。时间词代码为t,在语素的代码g前面置以t	秋/Tg 冬/tg 连/d 旱/a
u	助词	取英语助词auxiliary 的第2个字母,因a已用于形容词	工作/vn 的/u 政策/n
ud	结构助词		有/v 心/n 栽/v 得/ud 梧桐树/n
ug	时态助词		你/r 想/v 过/ug 没有/v
uj	结构助词的		迈向/v 充满/v 希望/n 的/uj 新/a 世纪/n
ul	时态助词了		完成/v 了/ ul
uv	结构助词地		满怀信心/l 地/uv 开创/v 新/a 的/u 业绩/n
uz	时态助词着		眼看/v 着/uz
v	动词		举行/v 老/a 干部/n 迎春/vn 团拜会/n
vd	副动词		强调/vd 指出/v
vg	动语素	动词性语素。动词代码为v。在语素的代码g前面置以V	做好/v 尊/vg 干/j 爱/v 兵/n 工作/vn
vn	名动词	指具有名词功能的动词。动词和名词的代码并在一起	股份制/n 这种/r 企业/n 组织/vn 形式/n ，/w
w	标点符号		生产/v 的/u ５Ｇ/nx 、/w ８Ｇ/nx 型/k 燃气/n 热水器/n
x	非语素字	非语素字只是一个符号，字母x通常用于代表未知数、符号	
y	语气词	取汉字“语”的声母	已经/d ３０/m 多/m 年/q 了/y 。/w
z	状态词	取汉字“状”的声母的前一个字母	势头/n 依然/z 强劲/a ； /w

英语词性对照表

Tag	Description	Example
CC	conjunction, coordinating	and, or, but
CD	cardinal number	five, three, 13%
DT	determiner	the, a, these
EX	existential there	there were six boys
FW	foreign word	mais
IN	conjunction, subordinating or preposition	of, on, before, unless
JJ	adjective	nice, easy
JJR	adjective, comparative	nicer, easier
JJS	adjective, superlative	nicest, easiest
LS	list item marker	
MD	verb, modal auxillary	may, should
NN	noun, singular or mass	tiger, chair, laughter
NNS	noun, plural	tigers, chairs, insects
NNP	noun, proper singular	Germany, God, Alice
NNPS	noun, proper plural	we met two Christmases ago
PDT	predeterminer	both his children
POS	possessive ending	's
PRP	pronoun, personal	me, you, it
PRP$	pronoun, possessive	my, your, our
RB	adverb	extremely, loudly
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	adverb, particle	about, off, up
SYM	symbol	%
TO	infinitival to	what to do?
UH	interjection	oh, oops, gosh
VB	verb, base form	think
VBZ	verb, 3rd person singular present	she thinks
VBP	verb, non-3rd person singular present	I think
VBD	verb, past tense	they thought
VBN	verb, past participle	a sunken ship
VBG	verb, gerund or present participle	thinking is fun
WDT	wh-determiner	which, whatever, whichever
WP	wh-pronoun, personal	what, who, whom
WP$	wh-pronoun, possessive	whose, whosever
WRB	wh-adverb	where, when
.	punctuation mark, sentence closer	.;?*
,	punctuation mark, comma	,
:	punctuation mark, colon	:
(	contextual separator, left paren	(
)	contextual separator, right paren	)

词性标注

中科院计算所分词系统

词性标注

应用

- 统计自然语言处理基础
- 句法分析的基础
- 辅助词义消歧
- 机器翻译
- and many more…

词性标注

词性标注歧义

- 一个词具有两个或者两个以上的词性
- 英文 Brown 语料库中，10.4% 的词是兼类词
  - the back door
  - on my back
  - promise to back the bill
- 汉语兼类词
  - 把门锁上 买了一把锁
  - 他研究与动物相关的研究工作
- 对兼类词消歧——词性标注的任务

词性标注

词性标注的信息来源是什么?

- 词汇信息

  当前词本身提供了关于标注的信息

- 句法结构信息

  考虑在当前词上下文中的词的词性

词性标注的性能指标

- 目前准确率大约在 97% 左右
- Baseline也可以达到 90%

  Baseline算法:

  对每一个词用它的最高频的词性进行标注

  未登录词全部标为名词

马尔可夫模型标注器

原理

- W —— 标注序列，T ——标注集，O —— 词集
- 马尔可夫模型标注器：假定一个词的词性只依赖于前一个词的词性（有限历史），而且这个依赖不随时间变化（时间不变）。

$$\underset{t_{1,\ldots,n}}{\arg\max} P(t_{1,\ldots,n} \mid w_{1,\ldots,n})$$

应用贝叶斯规则

$$= \underset{t_{1,n}}{\arg\max} \frac{P(w_{1,n} \mid t_{1,n}) P(t_{1,n})}{P(w_{1,n})}$$

$$= \underset{t_{1,n}}{\arg\max} P(w_{1,n} \mid t_{1,n}) P(t_{1,n})$$

马尔可夫模型标注器

原理

$$P(w_{1,n}\mid t_{1,n})P(t_{1,n})=\prod_{i=1}^{n}P(w_i\mid t_{1,n})$$

词的相互独立性

$$\times P(t_n\mid t_{1,n-1})\times P(t_{n-1}\mid t_{1,n-2})\times\ldots\times P(t_2\mid t_1)$$

一个词的词形只依赖于它自身的词性

$$=\prod_{i=1}^{n}P(w_i\mid t_i)$$

有限历史

$$\times P(t_n\mid t_{n-1})\times P(t_{n-1}\mid t_{n-2})\times\ldots\times P(t_2\mid t_1)$$

$$=\prod_{i=1}^{n}[P(w_i\mid t_i)\times P(t_i\mid t_{i-1})]$$

最终一个句子的最优标注序列公式为：

$$\hat{t}_{1,n}=\underset{t_{1,n}}{\operatorname{argmax}}P(t_{1,n}\mid w_{1,n})=\prod_{i=1}^{n}P(w_i\mid t_i)P(t_i\mid t_{i-1})$$

马尔可夫模型标注器

训练

- 有一个大的带标训练集
- 最大似然估计

$$\forall t^j \in tag, t^k \in tag, P(t^k \mid t^j) = \frac{C(t^j, t^k)}{C(t^j)}$$

$$\forall t^j \in tag, w^l \in word, P(w^l \mid t^j) = \frac{C(w^l, t^k)}{C(t^j)}$$

C(...) 是出现次数

课程提纲

一、词性标注（平滑技术）

二、语言模型

平滑处理

概念

- 为什么要做平滑处理?

平滑处理

概念

- 为什么要做平滑处理?

  处理数据矩阵的稀疏问题;

平滑处理

概念

- 为什么要做平滑处理?

  处理数据矩阵的稀疏问题;

- 为什么会稀疏?

平滑处理

概念

- 为什么要做平滑处理?

  处理数据矩阵的稀疏问题;

- 为什么会稀疏?

  计算实例的概率时，如果某个变量在观察样本库（训练集）中没有出现过。

拉普拉斯平滑

概念

- Laplace smoothing 或者 Additive smoothing
- N 个样本中，样本 i 的频率是 x_i，得到频率观察 $\mathrm{X} = \langle x_1, x_2, \ldots, x_i, \ldots, x_d \rangle$

$$p_{i,\ \text{empirical}} = \frac{x_i}{N}$$

拉普拉斯平滑

概念

- Laplace smoothing 或者 Additive smoothing
- N 个样本中，样本 i 的频率是 x_i，得到频率观察 $\mathrm{X} = \langle x_1, x_2, \ldots, x_i, \ldots, x_d \rangle$

$$p_{i,\ \text{empirical}} = \frac{x_i}{N}$$

$$p_{i,\ \alpha\text{-smoothed}} = \frac{x_i + \alpha}{N + \alpha d},$$

图灵估计

概念

- Good–Turing frequency estimation、Good-Turing 平滑法
- I. J. Good 1953 年提出。
- 基本思想：用观察计数较高的变量重新估计概率量的大小，并把它指派给那些具有零计数或者较低计数的变量。

图灵估计

例子

- 假设钓鱼过程中抓到了 18 条鱼，种类如下：

  10 carp, 3 perch, 2 whitefish, 1 trout, 1 salmon, 1 eel

- 那么，下一条鱼是 trout 的概率是多少？应该是 $\frac{1}{18}$
- 那么，下一条鱼是新品种的概率是多少？

  不考虑其他 0

  根据只出现一次的情况来估计 $\frac{3}{18}$

- 在此基础上，下一条鱼是 trout 的概率是多少？肯定小于 $\frac{1}{18}$
- 根据 Good Turing，对每一个计数 r，做一个调整，变为 r^*

$$r^* = (r+1)\frac{n_{r+1}}{n_r}$$

- 所以， $c = 1$时，

$$C^*(trout) = 2\times\frac{1}{3}; \qquad P^*(trout) = \frac{\frac{2}{3}}{18} = \frac{1}{27}$$

课程提纲

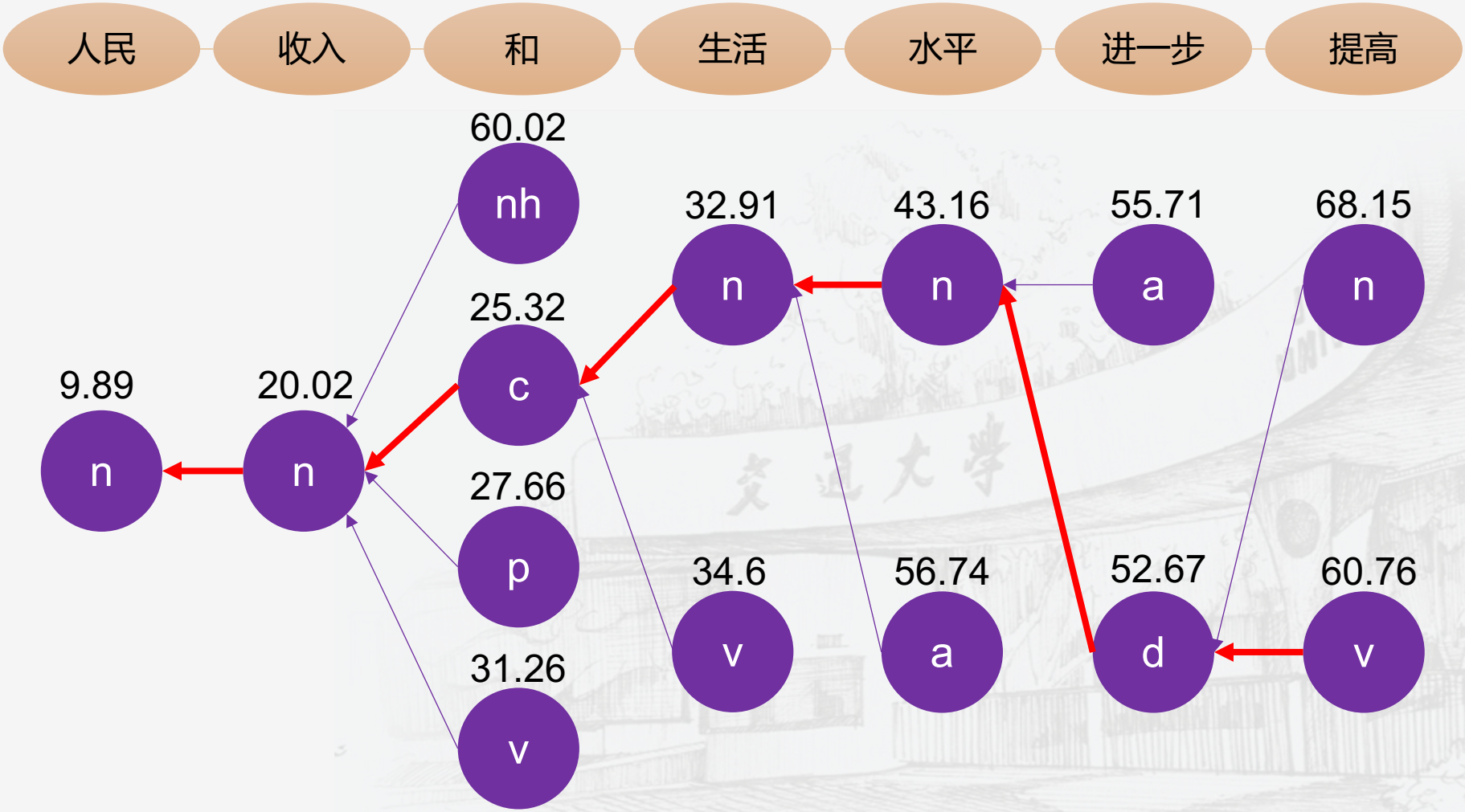
一、词性标注（继续）

二、语言模型

马尔可夫模型标注器

预测

- 如何求全局最优解



马尔可夫模型标注器

算法变形

- 未登录词
  - 考虑所有词性
  - 只考虑开放类词性
    - Uniform（平均分配概率）
    - Unigram（考虑每个词性独立出现的概率）
  - 根据未登录词的前缀和后缀猜测其词性
- 平滑
  - 平滑的原因：数据稀疏！
  - 收集更多的数据
  - 平滑
    - 估计在训练样本中没有出现情况的概率，降低已出现情况的概率来“分给”未出现的情况

马尔可夫模型标注器

注意

- 在训练时，我们能够观察到马尔可夫模型的状态，但是在标注时我们只能观察到词。所以我们说在马尔可夫模型标注时，实际上使用的是一个混合的方法。
- 在训练时构造VMMs，但是在标注时把他们当作HMMs。

为什么不直接称其为隐马尔科夫标注器？

- 隐马尔可夫标注器标注过程和马尔可夫标注器相同。
- 两者的区别在于怎样训练模型。
- 如果没有足够的训练语料，可以使用HMM来学习标注序列（使用前向后向算法）

基于转换的词性标注

基本思想

- 正确结果是通过不断修正错误得到的。
- 修正错误的过程是有迹可寻的。
- 让计算机学习修正错误的过程，这个过程可以用转换规则形式记录下来，然后用学习得到的转换规则进行词性标注。

基于转换的词性标注

基本思想

- 正确结果是通过不断修正错误得到的。
- 修正错误的过程是有迹可寻的。
- 让计算机学习修正错误的过程，这个过程可以用转换规则形式记录下来，然后用学习得到的转换规则进行词性标注。

方法评价

- 效果很好，在很多NLP问题上表现很好。
- 可以扩展到无监督学习中。

转换规则的形式

转换规则的组成

- 改写规则：将一个词性转换成另一个词性
- 激活环境：激发改写规则的条件

转换规则的形式

转换规则的组成

- 改写规则：将一个词性转换成另一个词性
- 激活环境：激发改写规则的条件

转换规则(T1)

- 改写规则：将一个词性从动词(v)改为名词(n)
- 激活环境：该词左边第一个词的词性是量词(q)

  第二个词的词性是数词(m)；

  举例：

  他/r 做/v 了/u 一/m 个/q 报告/v

  他/r 做/v 了/u 一/m 个/q 报告/n

转换规则的形式

转换规则的组成

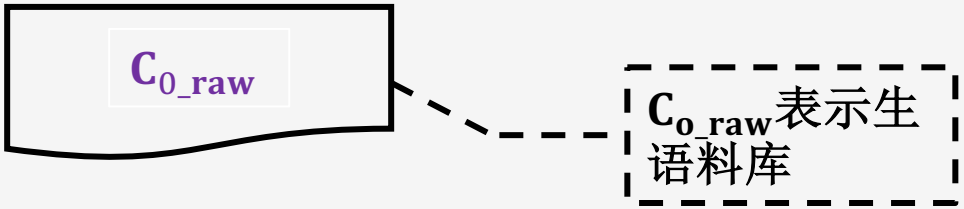
- 改写规则：将一个词性转换成另一个词性
- 激活环境：激发改写规则的条件

转换规则(T1)

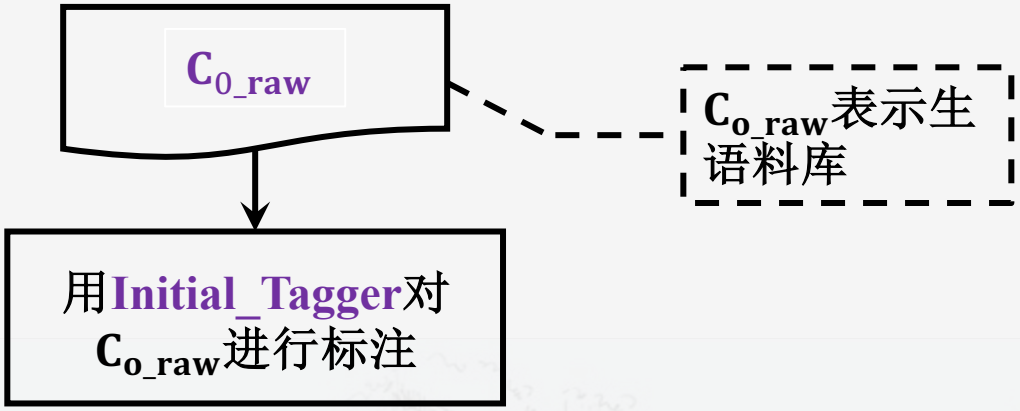
- 改写规则：将一个词性从动词(v)改为名词(n)
- 激活环境：该词左边第一个词的词性是量词(q)
  第二个词的词性是数词(m)；
  举例：
  他/r 做/v 了/u 一/m 个/q 报告/v
  他/r 做/v 了/u 一/m 个/q 报告/n

自然科学和社会科学的一个链接

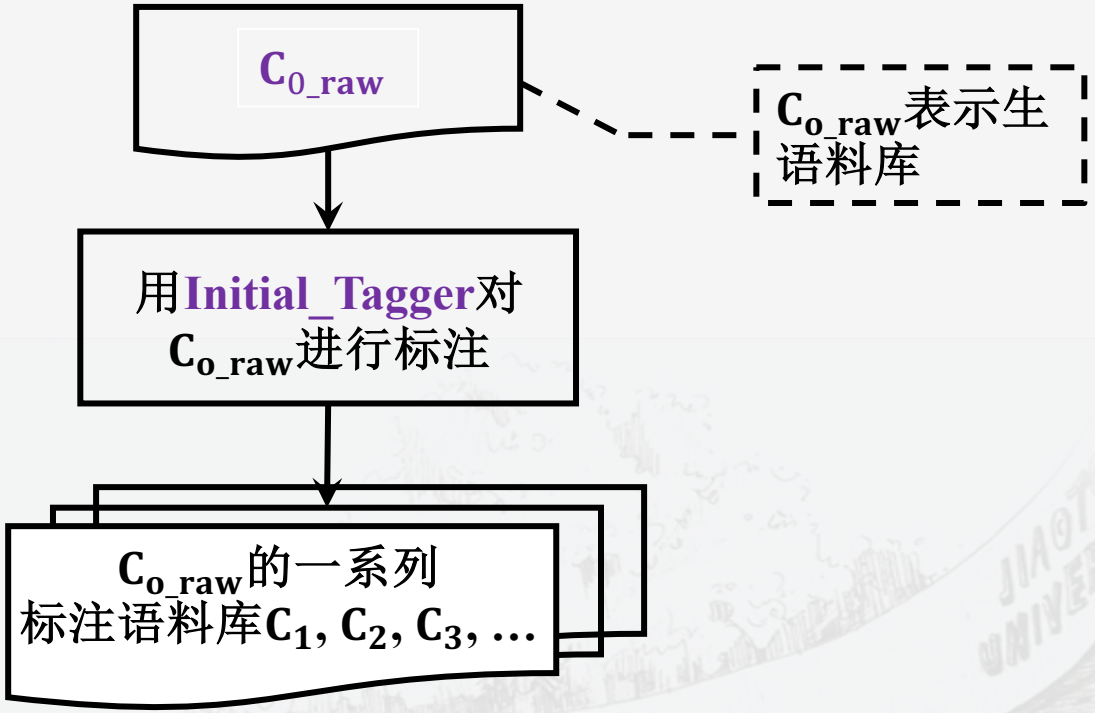
转换规则的学习流程



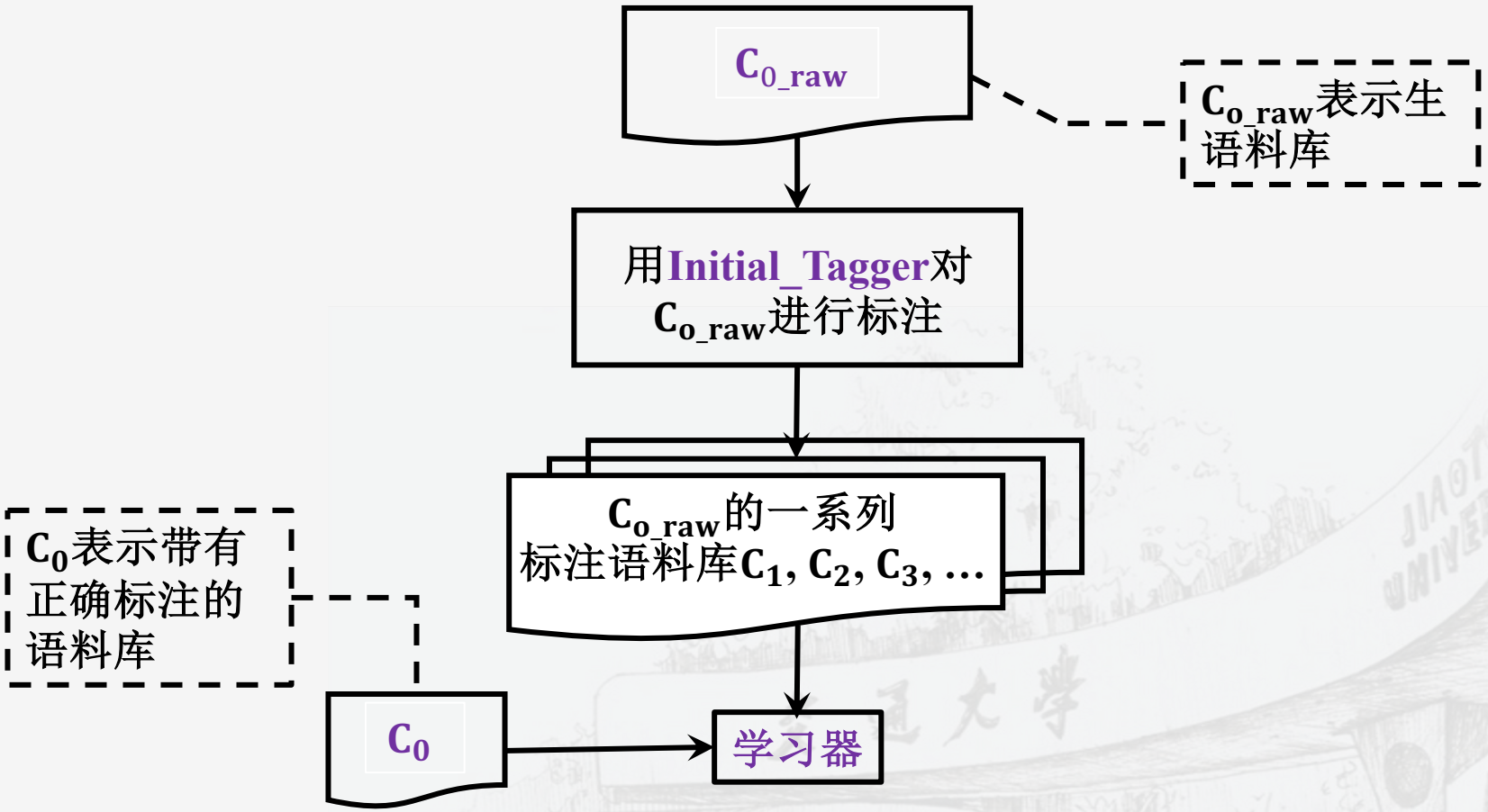
转换规则的学习流程



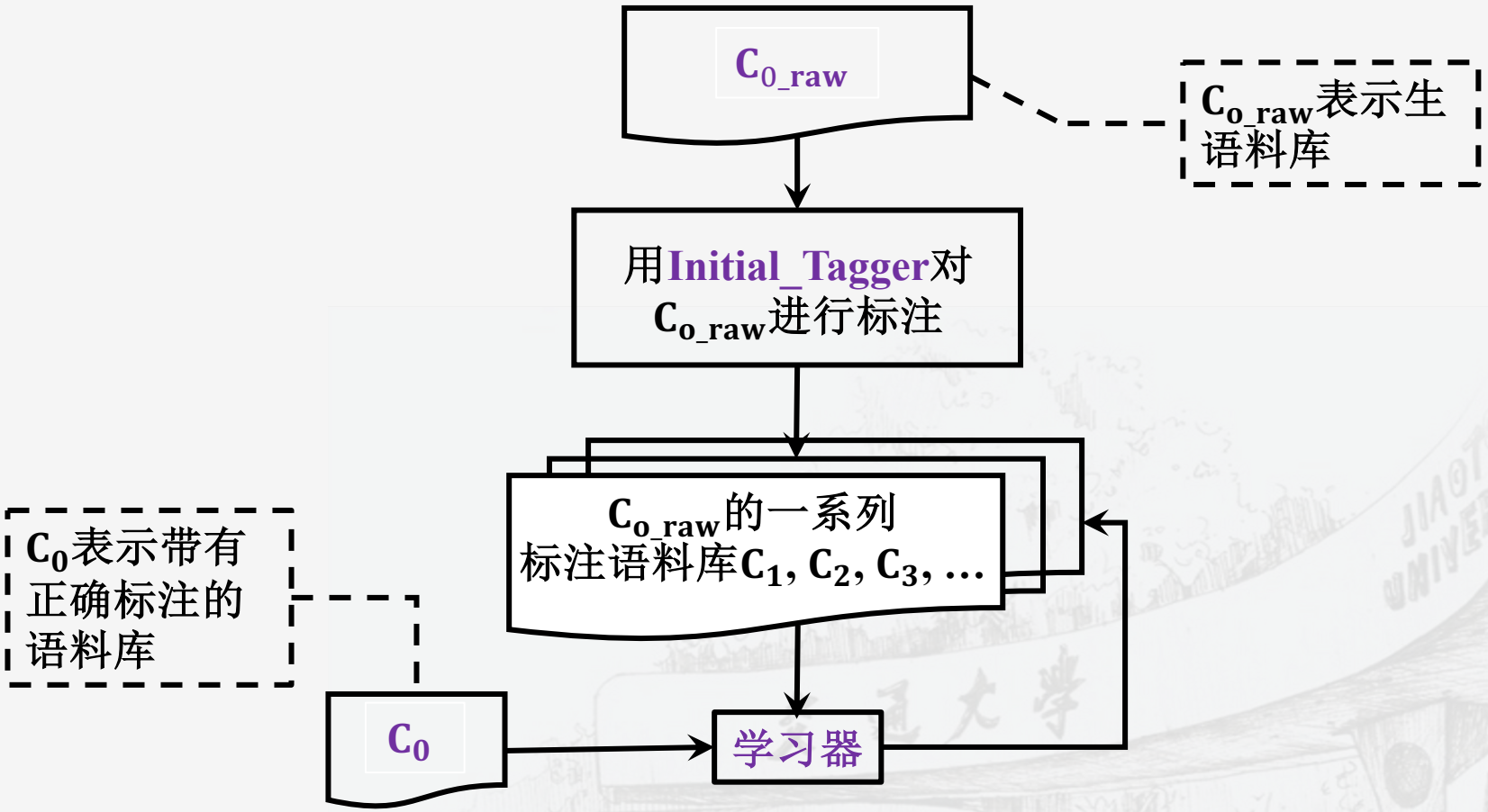
转换规则的学习流程


C_{0_raw}
C_{o_raw}表示生语料库
用Initial_Tagger对C_{o_raw}进行标注
C_{o_raw}的一系列标注语料库$C_1, C_2, C_3, \ldots$

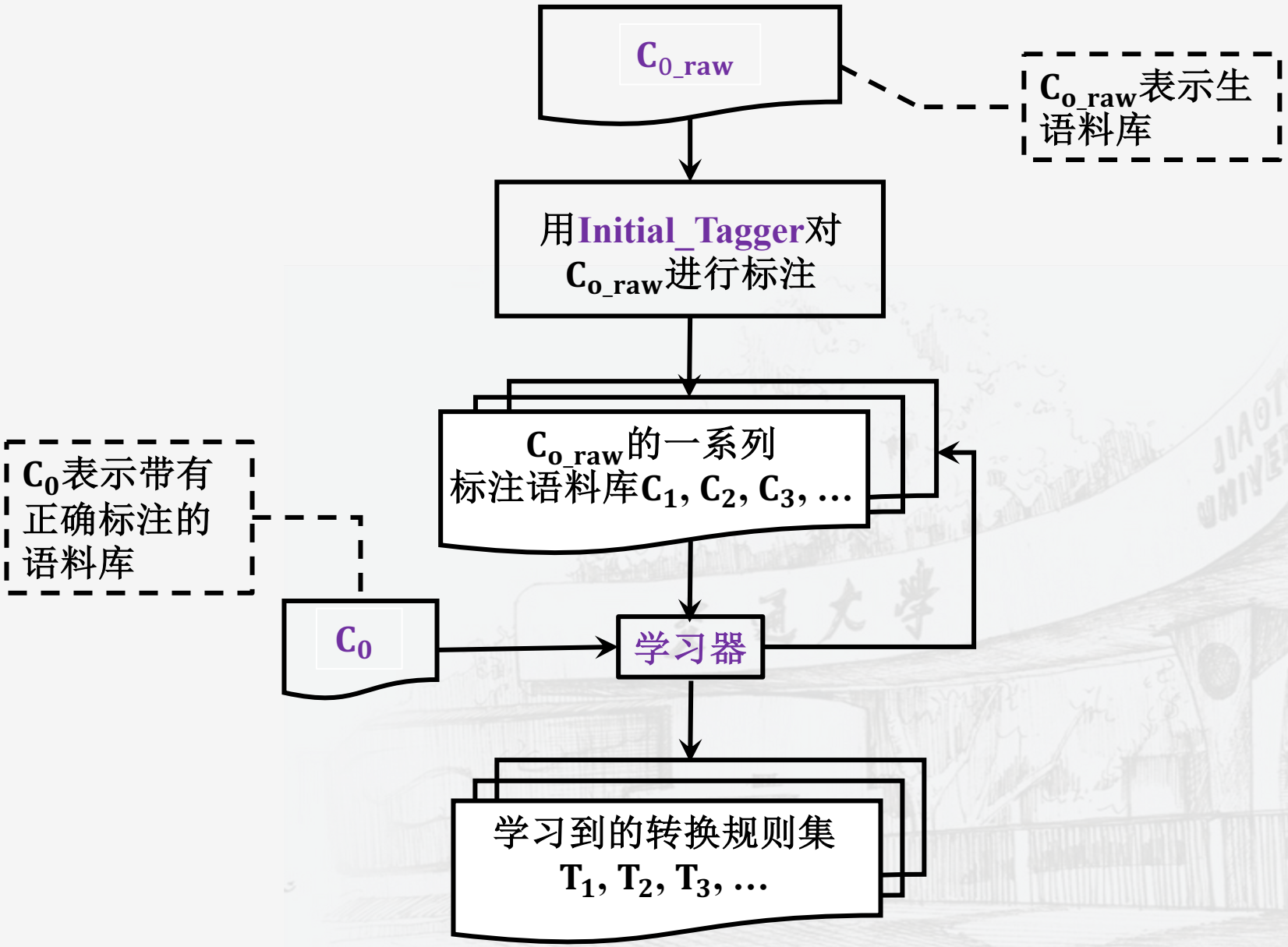
转换规则的学习流程


C_{0_raw}
C_{o_raw}表示生语料库
用Initial_Tagger对C_{o_raw}进行标注
C_{o_raw}的一系列标注语料库$C_1, C_2, C_3, \ldots$
C_0表示带有正确标注的语料库
C_0
学习器

转换规则的学习流程


C_{0_raw}
C_{o_raw}表示生语料库
用Initial_Tagger对C_{o_raw}进行标注
C_{o_raw}的一系列标注语料库$C_1, C_2, C_3, \ldots$
C_0表示带有正确标注的语料库
C_0
学习器

转换规则的学习流程


C_{0_raw}
C_{o_raw}表示生语料库
用Initial_Tagger对C_{o_raw}进行标注
C_{o_raw}的一系列标注语料库$C_1, C_2, C_3, \ldots$
C_0表示带有正确标注的语料库
C_0
学习器
学习到的转换规则集
$T_1, T_2, T_3, \ldots$

转换规则学习器算法

1. 首先用初始标注器对C_{0_raw}进行标注，得到带有词性标记的语料C_i (i =1);

转换规则学习器算法

1. 首先用初始标注器对C_{0_raw}进行标注，得到带有词性标记的语料C_i (i =1);
2. 将C_i跟正确的语料标注结果C_0比较，可以得到C_i中总的词性标注错误数;

转换规则学习器算法

1. 首先用初始标注器对C_{0_raw}进行标注，得到带有词性标记的语料C_i (i =1);
2. 将C_i跟正确的语料标注结果C_0比较，可以得到C_i中总的词性标注错误数;
3. 依次从候选规则中取出一条规则T_m(m=1,2...)，每用一条规则对C_i中的词性标注结果进行一次修改，就会得到一个新版本的语料库，不妨记做C_i^m(m=1,2,3...)，将每个C_i^m跟C_0比较，可计算出每个C_i^m中的词性标注错误数。假定其中错误数最少的那个是C_i^j(可预期C_i^j中的错误数一定少于C_i中的错数)，产生它的规则T_j就是这次学习得到的转换规则；此时C_i^j 成为新的待修改语料库，即$C_i^j = C_i^j$

转换规则学习器算法

1. 首先用初始标注器对C_{0_raw}进行标注，得到带有词性标记的语料C_i (i =1);
2. 将C_i跟正确的语料标注结果C_0比较，可以得到C_i中总的词性标注错误数;
3. 依次从候选规则中取出一条规则T_m(m=1,2...)，每用一条规则对C_i中的词性标注结果进行一次修改，就会得到一个新版本的语料库，不妨记做C_i^m(m=1,2,3...)，将每个C_i^m跟C_0比较，可计算出每个C_i^m中的词性标注错误数。假定其中错误数最少的那个是C_i^j(可预期C_i^j中的错误数一定少于C_i中的错数)，产生它的规则T_j就是这次学习得到的转换规则；此时C_i^j 成为新的待修改语料库，即$C_i^j = C_i^j$
4. 重复第3步的操作，得到一系列的标注语料库C_2^k, C_3^l, C_4^m,...后一个语料库中的标注错误数都少于前一个中的错误数，每一次都学习到一条令错误数降低最多的转换规则。直至运用所有规则后，都不能降低错误数，学习过程结束。这时得到一个有序的转换规则集合{T_a, T_b, T_c,...}

西安交通大学
XI'AN JIAOTONG UNIVERSITY

Q & A