



词性标注

刘秉权

哈工大智能技术与自然语言处理研究室

新技术楼612房间

liubq@hit.edu.cn



主要内容

- 简介
- 词性分类体系
- 标注中的信息源
- 主要标注方法
- 中文未登录词标注的词语特征



词性标注(POS(Part-of-Speech) Tagging)简介

- 任务：为句子中的每个词标上一个合适的词性

The-AT representative-NN put-VBD chairs-NNS on-IN the-AT table-NN.

The-AT representative-JJ put-NN chairs-VBZ on-IN the-AT table-NN.

- 词性指作为划分词类的根据的词的特点。
- 标注是一种有限语法消歧问题
- 目前最高准确率：96%-98%
- 应用
 - 信息抽取
 - 名词短语识别
 - 浅层句法分析
 - 问题回答



应用示例：

单纯依赖频率的搭配发现

$C(w^1 w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a



应用示例： 搭配过滤器的词性标记模式

Tag Pattern	Example
A N	<i>linear function</i>
N N	<i>regression coefficients</i>
A A N	<i>Gaussian random variable</i>
A N N	<i>cumulative distribution function</i>
N A N	<i>mean squared error</i>
N N N	<i>class probability function</i>
N P N	<i>degrees of freedom</i>



应用示例： 加词性过滤器的搭配发现

$C(w^1 w^2)$	w^1	w^2	Tag Pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N



中文词性标注实例

原文： 这件事情在理论界、经济界引起了很大反响。

分词后： 这 件 事 情 在 理 论 界 、 经 济 界 引 起 了 很 大 反
响 。

词性标注： 这/r 件/q 事情/n 在/p 理论界/n 、/w 经济
界/n 引起/v 了/u 很/d 大/a 反响/n 。/w



词性分类体系

■ 依据

- 形态标准
- 意义标准
- 分布标准--根据词在句法结构里所担当的语法功能分类，适合汉语的分类

■ 确定原则

- **标准性**：尽量采纳当前已经成为或正在成为词性标准的分类体系和标记符号
- 兼容性：尽量使标注集的表示与已经存在的标注集可以相互进行转化
- 扩展性：对未解决的遗留问题或是未来可能的技术发展方向充分加以考虑，以便加以扩充和修改



语言学界的分类

- 名词、时间词、处所词、方位词、动词、形容词、状态词、区别词、数词、量词、代词、介词、副词、连词、助词、语气词、象声词、叹词、前缀、后缀、成语、简称、习用语等



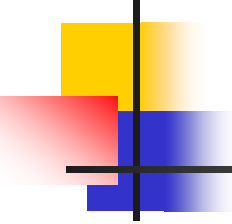
北京大学的汉语词性标注集

词性标记	词性	词性标记	词性	词性标记	词性
n	名词	z	状态词	h	前接成分
t	时间词	b	区别词	k	后接成分
s	处所词	d	副词	g	语素
f	方位词	p	介词	x	非语素字
m	数词	c	连词	i	成语
q	量词	u	助词	l	习用语
r	代词	y	语气词	j	简称略语
v	动词	o	拟声词	w	标点符号
a	形容词	e	叹词		



英语标注中常用的一些词性

Tag	Part Of Speech	Tag	Part Of Speech
AT	article	RB	adverb
BEZ	the word <i>is</i>	RBR	comparative adverb
IN	preposition	TO	the word <i>to</i>
JJ	adjective	VB	verb, base form
JJR	comparative adjective	VBD	verb, past tense
MD	modal	VBG	verb, present participle, gerund
NN	singular or mass noun	VBN	verb, past participle
NNP	singular proper noun	VBP	verb, non-3rd person singular present
NNS	plural noun	VBZ	verb, 3rd singular present
PERIOD	. : ? !	WDT	<i>wh</i> - determiner (<i>what, which</i>)
PN	personal pronoun		



■ 问题：词性由什么来确定？



标注中的信息源

- 邻近上下文中的其他词的标注
 - 很多词性序列很常见
 - 某些词性序列基本不可能
- 词本身提供的信息
 - Dumb标注器：简单地把最常用的标注分配给每个词，达90%准确率-基准性能
- 现代标注器都结合使用了结构语段信息和词汇信息



主要方法

- 马尔可夫模型标注器
- 隐马尔可夫模型标注器
- 基于转换的标注



马尔可夫模型标注器

- 将文本中的标记序列看成一条马尔可夫链

Limited horizon. $P(X_{i+1} = t^j | X_1, \dots, X_i) = P(X_{i+1} = t^j | X_i)$

Time invariant (stationary). $P(X_{i+1} = t^j | X_i) = P(X_2 = t^j | X_1)$



符号标记

w_i	the word at position i in the corpus
t_i	the tag of w_i
$w_{i,i+m}$	the words occurring at positions i through $i + m$ (alternative notations: $w_i \cdots w_{i+m}$, w_i, \dots, w_{i+m} , $w_{i(i+m)}$)
$t_{i,i+m}$	the tags $t_i \cdots t_{i+m}$ for $w_i \cdots w_{i+m}$
w^l	the l^{th} word in the lexicon
t^j	the j^{th} tag in the tag set
$C(w^l)$	the number of occurrences of w^l in the training set
$C(t^j)$	the number of occurrences of t^j in the training set
$C(t^j, t^k)$	the number of occurrences of t^j followed by t^k
$C(w^l : t^j)$	the number of occurrences of w^l that are tagged as t^j
T	number of tags in tag set
W	number of words in the lexicon
n	sentence length



问题求解

$$\begin{aligned}\arg \max_{t_{1,n}} P(t_{1,n}|w_{1,n}) &= \arg \max_{t_{1,n}} \frac{P(w_{1,n}|t_{1,n})P(t_{1,n})}{P(w_{1,n})} \\ &= \arg \max_{t_{1,n}} P(w_{1,n}|t_{1,n})P(t_{1,n})\end{aligned}$$



表达式简化

- 两个假设
 - 词语之间独立
 - 词语的出现只依赖于它本身的标注

$$\begin{aligned}\underline{P(w_{1,n}|t_{1,n})}P(t_{1,n}) &= \prod_{i=1}^n P(w_i|t_{1,n}) \\ &\quad \times P(t_n|t_{1,n-1}) \times P(t_{n-1}|t_{1,n-2}) \times \cdots \times P(t_2|t_1) \\ &= \prod_{i=1}^n P(w_i|t_i) \\ &\quad \times P(t_n|t_{n-1}) \times P(t_{n-1}|t_{n-2}) \times \cdots \times P(t_2|t_1) \\ &= \prod_{i=1}^n [P(w_i|t_i) \times P(t_i|t_{i-1})]\end{aligned}$$



确定一个句子的最优标注

$$\hat{t}_{1,n} = \arg \max_{t_{1,n}} P(t_{1,n} | w_{1,n}) = \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1})$$

$$P(t^k | t^j) = \frac{C(t^j, t^k)}{C(t^j)}$$

$$P(w^l | t^j) = \frac{C(w^l, t^j)}{C(t^j)}$$



模型训练算法

```
1 for all tags  $t^j$  do
2   for all tags  $t^k$  do
3      $P(t^k | t^j) := \frac{C(t^j, t^k)}{C(t^j)}$ 
4   end
5 end
6 for all tags  $t^j$  do
7   for all words  $w^l$  do
8      $P(w^l | t^j) := \frac{C(w^l, t^j)}{C(t^j)}$ 
9   end
10 end
```



Brown语料库中一些标记转移的计数

First tag	Second tag					
	AT	BEZ	IN	NN	VB	PERIOD
AT	0	0	0	48636	0	19
BEZ	1973	0	426	187	0	38
IN	43322	0	1325	17314	0	185
NN	1067	3720	42470	11773	614	21392
VB	6072	42	4758	1476	129	1522
PERIOD	8016	75	4656	1329	954	0



Brown语料库中一些词和标记共现的计数

	AT	BEZ	IN	NN	VB	PERIOD
<i>bear</i>	0	0	10	0	43	0
<i>is</i>	0	10065	0	0	0	0
<i>move</i>	0	0	0	36	133	0
<i>on</i>	0	0	5484	0	0	0
<i>president</i>	0	0	0	382	0	0
<i>progress</i>	0	0	0	108	4	0
<i>the</i>	69016	0	0	0	0	0
<i>.</i>	0	0	0	0	0	48809



标注算法(Viterbi算法)

```
1 comment: Given: a sentence of length  $n$ 
2 comment: Initialization
3  $\delta_1(\text{PERIOD}) = 1.0$ 
4  $\delta_1(t) = 0.0$  for  $t \neq \text{PERIOD}$ 
5 comment: Induction
6 for  $i := 1$  to  $n$  step 1 do
7     for all tags  $t^j$  do
8          $\delta_{i+1}(t^j) := \max_{1 \leq k \leq T} [\delta_i(t^k) \times P(w_{i+1}|t^j) \times P(t^j|t^k)]$ 
9          $\psi_{i+1}(t^j) := \arg \max_{1 \leq k \leq T} [\delta_i(t^k) \times P(w_{i+1}|t^j) \times P(t^j|t^k)]$ 
10    end
11 end
12 comment: Termination and path-readout
13  $X_{n+1} = \arg \max_{1 \leq j \leq T} \delta_{n+1}(j)$ 
14 for  $j := n$  to 1 step - 1 do
15      $X_j = \psi_{j+1}(X_{j+1})$ 
16 end
17  $P(X_1, \dots, X_n) = \max_{1 \leq j \leq T} \delta_{n+1}(t^j)$ 
```



与隐马尔可夫模型区别

- 训练时构造了“显”马尔可夫模型
- 标注时当作了隐马尔可夫模型

模型参数A、 B 、 π 都可根据语料库来求出

t不知道，w知道



其他问题

- 未登录词
- 三元语法标注器
- 插值和可变记忆
- 平滑



隐马尔可夫标注器

- HMM标注过程与VMM相同
- 差别在于怎样训练模型
- HMM的初始化是关键所在



HMM的初始化

- 随机地初始化HMM的所有参数：使得标注难以约束
- 使用词典信息限制模型参数：如果对应的词语-标记对未在词典中列出，则将词语生成概率设为0
- 将词语聚集到词语等价类中，使得所有同一类的词语允许同样的标注

前面是基于概率

基于转换的标注学习

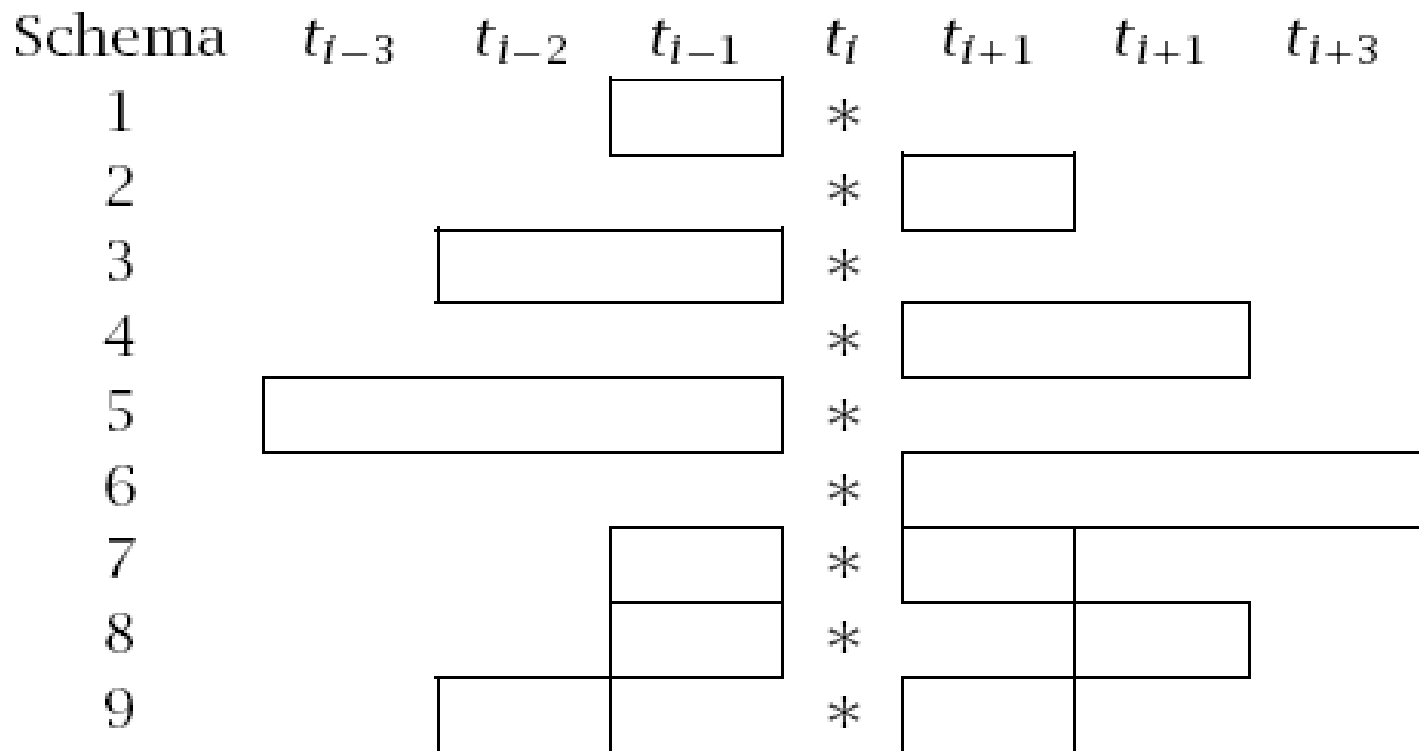
- 利用更大范围的词汇和语法结构规则
- 标注器需要的决策量比估计大量的马尔可夫模型参数要少一个数量级

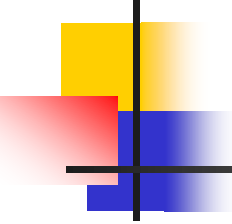


主要思想

- 给定一个标注好的语料库和词典
- 用最常用的标记标注训练语料中的每个词
- 构建一个转换的序列表，它将初始标注转化为接近正确的标注
- 使用转换序列表标注新的文本：
 - 初始化新的文本，用最常用的标记来标注
 - 应用转换

基于转换的标注器中的触发环境





学习到的一些转换例子

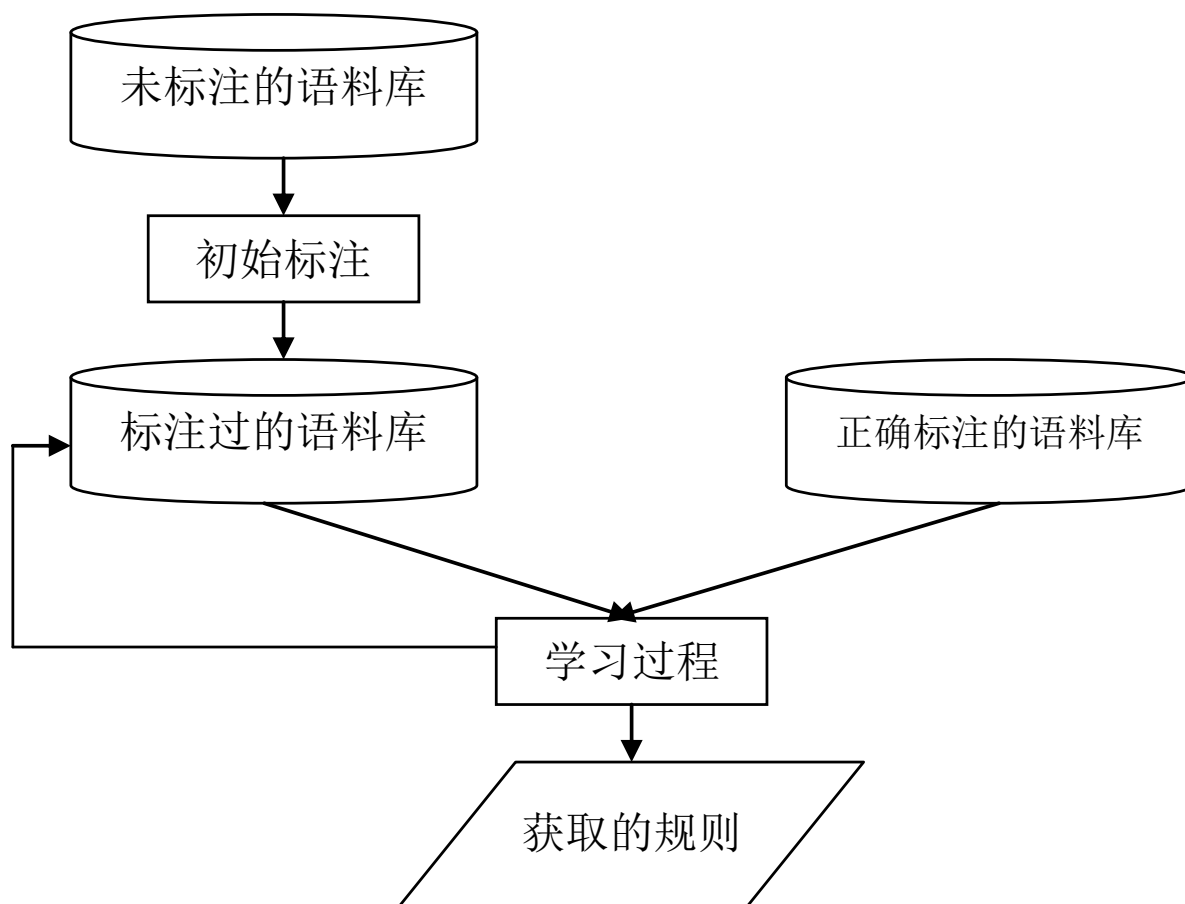
Source tag	Target tag	Triggering environment
NN	VB	previous tag is TO
VBP	VB	one of the previous three tags is MD
JJR	RBR	next tag is JJ
VBP	VB	one of the previous two words is <i>n't</i>



触发条件

- 词语触发
- 标记触发
- 联合触发
- 形态触发：处理未登录词

基于转换的标注学习示意图





基于转换的标注学习算法

```
1  $C_0$  := corpus with each word tagged with its most frequent tag
3 for  $k := 0$  step 1 do
4      $v$  := the transformation  $u_l$  that minimizes  $E(u_l(C_k))$ 
6     if ( $E(C_k) - E(v(C_k))$ )  $< \epsilon$  then break fi
7      $C_{k+1} := v(C_k)$ 
8      $\tau_{k+1} := v$ 
9 end
10 Output sequence:  $\tau_1, \dots, \tau_k$ 
```



分析

- 准确率：95%-96%
- 能将标注决策建立在更丰富的事件集合上
- 转换比概率标注中的转移和词语生成更容易修改
- 基于转换的学习也被应用于句法分析、介词短语附着、语义消歧上



其他标注方法

- 神经元网络
- 决策树
- K近邻方法
- 最大熵模型



标注准确率

- 影响标注性能的因素
 - 可以获得的训练数据量：通常越多越好
 - 标记集：标记集越大，潜在歧义越多，标注越困难
 - 训练语料库及词典与应用语料库的差别
 - 未登录词



概率标注器中常见的错误例子

Correct tag	Tagging error	Example
noun singular	adjective	<i>an <u>executive</u> order</i>
adjective	adverb	<i><u>more</u> important issues</i>
preposition	particle	<i>He ran <u>up</u> a big ...</i>
past tense	past participle	<i>loan <u>needed</u> to meet</i>
past participle	past tense	<i>loan <u>needed</u> to meet</i>



词性标注混乱矩阵的一部分

Correct Tags	Tags assigned by the tagger							
	DT	IN	JJ	NN	RB	RP	VB	VBG
DT	99.4	.3			.3			
IN	.4	97.5			1.5	.5		
JJ		.1	93.9	1.8	.9		.1	.4
NN			2.2	95.5			.2	.4
RB	.2	2.4	2.2	.6	93.2	1.2		
RP		24.7		1.1	12.6	61.5		
VB			.3	1.4			96.0	
VBG			2.5	4.4				93.0



中文未登录词标注的词语特征

- 后缀特征
- 部首特征
- 重叠特征



中文词语后缀特征

- 主要被用于识别地名、机构名或其他专有名词
 - 后缀为“市”,如北京市(地名,ns),多为名词;
 - 后缀为“化”,如市场化(动词,v),多为动词;
 - 后缀的组合为“部门”,如政府部门(名词,n),多为名词;
 - 后缀的组合为“委会”,如奥委会(简称,j)、特委会(简称,j)、村委会(简称,j),多为简称。



部首特征

- 某一部首下所列的一系列具体汉字几乎都与该部首有着意义上的联系
 - 列在“木”部的字如杨、柳、森、林等,都与“木”相关;
 - 列入“车”部的字如轮、轻、辑、轩等,都与“车”有关;
 - 列入“示”部的字如神、祖、禅、祀等,都与祭祀有关。
- 利用部首初步猜测词性
 - 言字旁说、记、论等,一般为动词;
 - 立刀旁刖、刮、判等,一般为动词;
 - 提土旁地、场、城等,一般为名词。
- 特征提取与组合
 - 通过Unicode编码提取
 - 当知道一个汉字的部首时,可以初步猜测该字的词性;
 - 但是当由该字组成一个词时,则不能简单地通过一个字来猜测该词的词性,需要一个词所含字的部首的组合。



重叠特征

- 去重后判断词性,原词语与去重后的词在词性上一般是相同的
 - 高高兴兴可以通过“高兴”来判断词性
 - 轻轻的可以根据“轻”来判断词性等
- 重叠词的提取
 - 当重叠词是类似于“高高兴兴”这种形式时,提取“高兴”
 - 当它是“轻轻地/的”这种形式时,则提取“轻”
 - 当它是“湛蓝湛蓝”这种形式时,则提取“湛蓝”

谢谢！

