词类自动标注

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

词的分类依据

- 形态标准

  *Words that function similarly with respect to the **affixes** they take (their **morphological properties**) are grouped into classes.*

- 分布标准

  *Words that function similarly with respect to what can occur **nearby** (their "**syntactic distributional properties**") are grouped into classes.*

- 意义标准(×)

  *While word classes do have tendencies toward **semantic coherence** (nouns do in fact often describe "people, places or things", and adjectives often describe properties), this is not necessarily the case, and in general **we don't use the semantic coherence as a definition criterion for part-of-speech**.*

英语中词的分类

- 英语词类
  - ***preposition***, ***determiner***, *pronoun*, ***conjunction***, *nouns, verbs, adjectives, adverbs, numeral, interjection*
- *closed class* and *open class*
  - *Closed classes are those that have relative fixed membership, in which new words are rarely coined.*
- *function word* and *content word*

英语中词的分类

- 词类的子类
  - noun
    - proper noun *eg. Beijing, IBM*
    - common noun
      - countable noun *eg. book, table*
      - mass noun *eg. communism, salt*
  - adverb
    - directional adverb *eg. downhill, home*
    - degree adverb *eg. somewhat, extremely, very*
    - manner adverb *eg. slowly, delicately*
    - temporal adverb *eg. yesterday, tomorrow*

汉语中词的分类

- 汉语中词的分类依据
  - 汉语缺乏形态，形态特征不能用作分类依据。
  - 概念相近的词，语法性质未必相同，意义不能作为分类依据。例：战争(名词)、战斗(动词)
  - 汉语中词分类的依据主要是词的分布特征，或者说主要依据词的语法功能。
- 词的语法功能主要指词在句法结构里所能占据的语法位置。
  - 词在句法结构中充当句法成分的能力
  - 词与某类词或某些词组合成短语的能力
- 虽然不能根据意义对词进行分类，但按照分布特征同属一类的词，意义上也常有共性。
  - 名词通常表示事物的名称、动词通常表示动作和行为、形容词表示事物的性质和状态。

汉语中词的分类

- 实词和虚词
  - 从功能上看，实词可以充当主语、谓语和宾语。虚词则不可以。
  - 从意义上看，实词有实在的意义，表示事物、动作、行为、变化、性质、状态、处所、时间等。虚词基本只起语法作用，本身多无实在意义。
  - 从数量上看，实词多为开放类，虚词多为封闭类。
- 体词和谓词
  - 实词通常可进一步分成体词和谓词。体词可以做主语和宾语。谓词主要做谓语。

汉语中词的分类

- 体词
  - 名词(1)、处所词(2)、方位词(3)、时间词(4)、区别词(5)、数词(6)、量词(7)、代词(8)
- 谓词
  - 动词(9)、形容词(10)
- 虚词
  - 副词(11)、介词(12)、连词(13)、助词(14)、语气词(15)
- 拟声词(16)、感叹词(17)

朱德熙，《语法讲义》，商务印书馆，1982

汉语中词的分类

- 为什么说一个词是形容词？
  - 可以用作主谓结构中的谓语，但不能带真宾语。
    - 例：长江比黄河**长**、**长**三公分
  - 可以受“很”这类程度副词修饰。例：很**长**、很**雄伟**、很**安静**
  - 可以作述补结构中的补语。例：洗**干净**、捆**结实**
  - 直接或加“地”后作状中结构中的状语。例：**迅速**提高
  - 直接或加“的”后作定中结构中的定语。例：**美丽**人生
  - 可以用“a + 不 + a” 的形式提问。例：**舒服**不**舒服**？
  - ……

俞士汶等，《现代汉语语法信息词典详解》,清华大学出版社，2003

汉语中词的分类

- 对汉语词类问题有兴趣，可进一步参考有关书籍。
- 由于汉语缺乏形态，词的类别不如英语等西方语言那样易于判别。汉语语言学家曾在汉语词类划分问题上有过不同意见，并经过长期争议，至今仍然存在多种看法。
- 利用计算机处理语言，词语的语法分类及其代码化不可缺少。面向信息处理用汉语词类体系的建立和大规模词语归类实践必须进行。

兼类问题

- 如果同一个词具有不同词类的语法功能，则认为这个词兼属不同的词类，简称兼类。
- 例一

  (1a) **共同**完成一些任务　　(1b) 我们的**共同**愿望
  (2a) **自动**控制这个开关　　(2b) 方便的**自动**步枪
  (3a) **定期**检查机器　　(3b) 一笔**定期**存款
  在(a)组中，是副词、在(b)组中是区别词。

- 例二

  (4a) 买了一束**花**　　(4b) **花**了很多时间
  (5a) 开了一个**会**　　(5b) **会**拉小提琴
  (6a) 桌子上有两封**信**　　(6b) **别**信他的话
  (7a) 选举他当**代表**　　(7b) 他**代表**我们发言
  在(a)组中是名词，在(b)组中是动词。

兼类问题

- English data, from Brown corpus:
  - 11.5 percent of the lexicon is ambiguous as to part-of-speech (types)
  - 40 percent of the words in the Brown corpus are ambiguous (tokens)
- Degree of ambiguity (No. tags per word)

1 tags	35340		
2-7 tags	4100	total:	39440
2 tags	3760		
3 tags	264		
4 tags	61		
5 tags	12		
6 tags	2		
7 tags	1		

兼类问题

- 《现代汉语语法信息词典》数据(1997年版)
  - 总词数 55191
  - 2-5 tags 1624 2.94%
  - 2 tags 1475 2.67%
  - 3 tags 126 0.23%
  - 4 tags 20 0.04%
  - 5 tags 3 0.01%
- 例：
  - 和 c-n-p-q-v
  - 光 a-d-n-v

英语词类标记集

- *Brown corpus tagset*
  - 87 *tags*
  - *Used for Brown Corpus* (1-million-word,1963-1964, Brown University)
  - TAGGIT program
- *Penn treebank tagset*
  - 45 *tags*
  - *Used for Penn treebank, Brown Corpus, WSJ Corpus*
  - *Brill tagger*
- *UCREL's C5 tagset*
  - 61 *tags*
  - *Used for British National Corpus (BNC)*
  - *Lancaster CLAWS tagger*

英语词类标记集

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	*+,%, &*
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Penn treebank POS tagset

(45 *tags*)

汉语词类标记集

- 北京大学《人民日报》语料库词类标记集
  - 规范2001
    - 39个词类标记
    - 用于标注《人民日报》语料库
    - 词类标记规范参见
      - http://www.icl.pku.edu.cn/icl_groups/corpus/coprus-annotation.htm
  - 规范2003
    - 扩充至106个词类标记
- 国家语委语用所词类标记集
  - ??个词类标记
  - 参见语委用用所《信息处理用现代汉语词类及词性标记集规范》
- 其它词类标记集

汉语词类标记集

标记	描述	标记	描述
Ag	形语素	ns	地名
a	形容词	nt	机构团体
ad	副形词	nz	其他专名
an	名形词	o	拟声词
b	区别词	p	介词
c	连词	q	量词
Dg	副语素	r	代词
d	副词	s	处所词
e	叹词	Tg	时语素
f	方位词	t	时间词
g	语素	u	助词
h	前接成分	Vg	动语素
i	成语	v	动词
j	简称略语	vd	副动词
k	后接成分	vn	名动词
l	习用语	w	标点符号
m	数词	x	非语素字
Ng	名语素	y	语气词
n	名词	z	状态词
nr	人名		

北大《人民日报》标注语料库词类标记集(39 tags)

为了处理真实语料，汉语词类标记集中通常包含一些非功能分类的标记，例如：成语、习用语、简称略语；也包含一些语素、前接成份、后接成份等比词小的标记。

词类自动标注

- 词类自动标注的任务
  - 判定自然语言句子中的每个词的词类并给每个词赋以词类标记。
  - 例如：
    - book that flight.
      book/VB that/DT flight/NN ./.
    - 这 份 特区 政府 的 报告 长 达 2 0 页 。
      这/r 份/q 特区/n 政府/n 的/u 报告/n 长/a 达/v 2 0/m 页/q 。/w
- 对于兼类词，词类标注程序应根据上下文确定兼类词在句子中最合适的词类标记。(难点所在)

词类自动标注

- 词类自动标注是深层语言分析的基础
  - 句法分析
- 词类标注程序判定依据
  - 要标注的词的不同词类的分布
    - can　MD-VB-NN　　(大部分情形是MD)
    - dumb tagger 统计每个兼类词的词类概率分布，并给每个词赋概率最大的词类。
      - 对英语而言，试验结果 90%
  - 上下文中其它词的词类信息
    - 英语中，词类串DT JJ NN比DT JJ VBZ更加可能。

词类自动标注

- 基本方法
  - 基于规则的词类标注
  - 基于统计的词类标注
  - 统计规则相结合的词类标注

基于规则的词类标注方法

- 早期的词类标注方法多为基于规则的方法
  - 70年代初 TAGGIT 标注程序
  - 约 3300 规则人工总结的规则
  - 标注Brown语料库（87 tags），准确率约77%
- 目前的基于规则的词类标注程序性能远远好于TAGGIT标注程序
- 基于规则的词类标注程序工作过程
  1. 查词典，给句中各词标记所有可能的词类标记。
  2. 应用规则，逐步删除错误的标记，最终只留下正确的标记。

基于规则的词类标注

- 以 EngCG tagger (1995)为例
- 查词典，给句中各词标记可能的词类标记以及特征信息(形态特征、次范畴化框架特征等)

Pavlov had shown that salivation …

↓

Pavlov	**PAVLOV N NOM SG PROPER**
had	**HAVE V PAST VFIN SVO**
	HAVE PCP2 SVO
shown	**SHOW PCP2 SVOO SVO SV**
that	ADV
	PRON DEM SG
	DET CENTRAL DEM SG
	CS
salivation	**N NOM SG**
. . .	

基于规则的词类标注

- 规则(constraint)示例

ADVERBIAL-THAT RULE
Given input: "that"
if
(+1 A/ADV/QUANT); /* *if next word is adj, adverb, or quantifier* */
(+2 SENT-LIM); /* *and following which is a sentence boundary,* */
(NOT -1 SVOC/A); /* *and the previous word is not a verb like* */
/* *'consider' which allows adjs as object complements* */
then eliminate non-ADV tags
else eliminate ADV tag

规则用以删除和上下文环境不相容的标记。

*it isn't **that** odd.*

*I consider **that** odd.*

基于隐马尔科夫模型的词类标注

- HMM状态集　　词类标记集
- HMM输出符号集　　词表
- 如何根据观察到的词串(句子)，求解最可能的词类标记序列(状态转换序列)。　　维特比算法
- 模型参数
  - $p(t_i|t_{i-1})$　　词类转移概率
  - $p(w_i|t_i)$　　词类t_i生成词w_i的概率
  - $p(t)$　　词类t出现在句首的概率

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}}\, P(t_1^n \mid w_1^n) = \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} p(w_i \mid t_i) p(t_i \mid t_{i-1})$$

基于隐马尔科夫模型的词类标注

- 参数学习
  - 多采用有指导的学习方法
  - 需要预先准备带词类标记的语料库
    - 例如，1998年1月《人民日报》标注语料库
  - 也可以采用无指导学习，例如用Baum-Welch算法
- 最大似然估计

$$p(t_i \mid t_{i-1}) = \frac{c(t_{i-1}, t_i)}{c(t_{i-1})}$$

特殊标记<*BOS*> <*EOS*>

$$p(w_i \mid t_i) = \frac{c(t_i, w_i)}{c(t_i)}$$

汉语词类标注实例

- 1998年1月《人民日报》标注语料
- 作为动词的“报告”(30次)
  1... 5 3 岁的福塞特向总部**报告**说，负责热气球...
  2...将刘青山、张子善的严重犯罪事实**报告**党中央，...
  3...有关矿产资源情况，要每周向中央主要领导**报告**。
- 作为名词的“报告”(200次)
  1...在党的十五大**报告**中，江主席再次郑重地...
  2...**报告**认为，虽然日本政府为减少限制性贸易...
  3...国际金融协会发表资金流动**报告**...

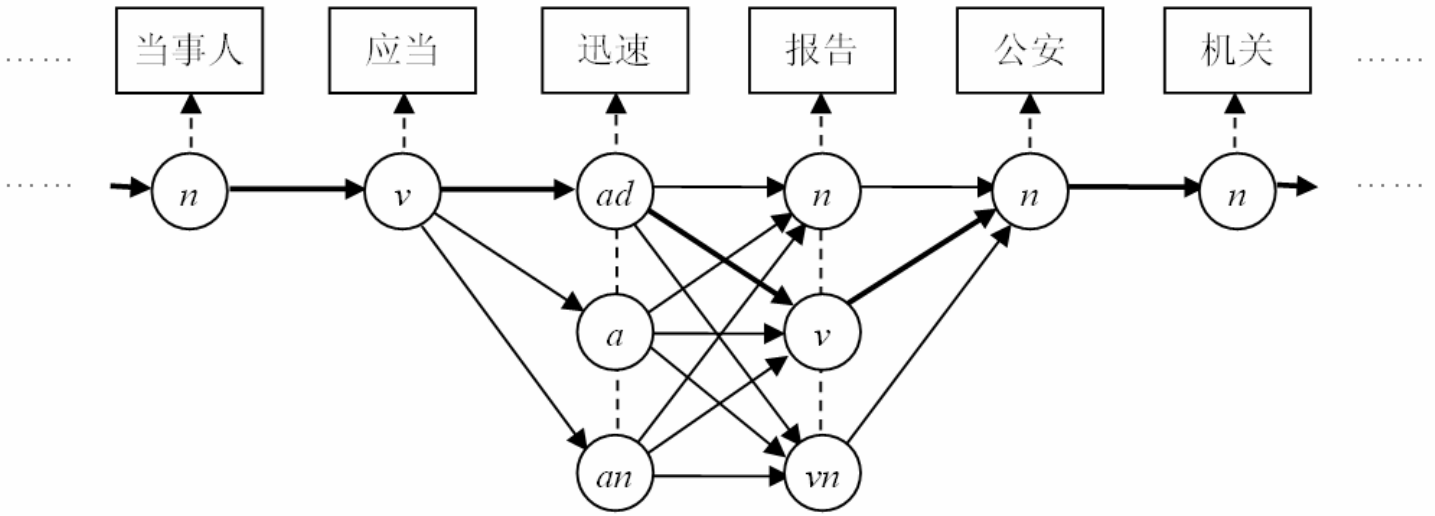
? 发生交通事故时，当事人应当迅速**报告**公安机关，听候处理...

汉语词类标注实例

$c(t_i,t_{i+1})$	…	a	ad	an	n	v	vn	…	Σ
…	…		…	…	…	…	…	…	…
a		800	8	127	10923	942	2267		34473
ad	…	76	34	0	3	5533	2	…	5933
an	…	10	5	47	238	257	218	…	2837
n	…	4047	1273	440	42491	32933	12508	…	312263
v	…	6924	855	735	42671	27142	4735	…	229776
vn	…	284	113	54	16021	2677	3165	…	42734
…	…		…						

$c(w_i,t_i)$	…	a	ad	an	n	v	vn	…	Σ
…	…	…	…	…	…	…	…	…	…
当事人	…	0	0	0	25	0	0	…	25
应当	…	0	0	0	0	340	0	…	340
迅速	…	50	116	1	0	0	0	…	167
报告	…	0	0	0	200	30	4	…	234
公安	…	0	0	0	188	0	0	…	188
机关	…	0	0	0	354	0	0	…	354
…	…	…	…		…	…	..	…	…

汉语词类标注实例



P(… n v ad n n n …|… 当事人 应当 迅速 报告 公安 机关 …)

=…×p(当事人|n)×p(v|n)×p(应当|v)×p(ad|v)×p(迅速|ad)×p(n|ad)

×p(报告|n)×p(n|n)×p(公安|n)×p(n|n)×p(机关|n)×…

汉语词类标注实例

$p(t_{i+1}\|t_i)$	…	a	ad	an	n	v	vn	…	Σ
…	…		…	…	…	…	…	…	…
a		0.0232065676	0.0002320657	0.0036840426	0.3168566704	0.0273257332	0.0657616105		
ad	…	0.0128097084	0.0057306590	0	0.0005056464	0.9325804820	0.0003370976	…	1
an	…	0.0035248502	0.0017624251	0.0165667959	0.0838914346	0.0905886500	0.0768417342	…	1
n	…	0.0129602290	0.0040766918	0.0014090686	0.1360743988	0.1054655851	0.0400559785	…	1
v	…	0.0301336954	0.0037210152	0.0031987675	0.1857069494	0.1181237379	0.0206070260	…	1
vn	…	0.0066457622	0.0026442645	0.0012636308	0.3749005476	0.0626433285	0.0740628071	…	1
…	…		…		…	…	…	…	…

汉语词类标注实例

$p(w_i\|t_i)$	…	a	ad	an	n	v	vn	…
…	…	…	…	…	…	…	…	…
当事人	…	0	0	0	0.0000800607	0	0	…
应当	…	0	0	0	0	0.0014797019	0	…
迅速	…	0.0014504105	0.0195516602	0.0003524850	0	0	0	…
报告	…	0	0	0	0.0006404857	0.0001305619	0.0000936023	…
公安	…	0	0	0	0.0006020566	0	0	…
机关	…	0	0	0	0.0011336598	0	0	…
…	…	…	…	…	…	…	..	…
Σ	…	1	1	1	1	1	1	…

汉语词类标注实例

T	$\prod p(w_i\|t_i)\,p(t_i\|t_{i-1})$	$P(T\|$...应当 迅速 报告 公安...$)$
...v ad n n...	...p(ad|v)×p(迅速|ad)×p(n|ad)×p(报告|n)×p(n|n)...	3.2061059e-12
...v ad v n...	**...p(ad|v)×p(迅速|ad)×p(v|ad)×p(报告|v)×p(n|v)...**	**1.64503834e-9**
...v ad vn n...	...p(ad|v)×p(迅速|ad)×p(vn|ad)×p(报告|vn)×p(n|vn)...	8.6060396e-13
...v a n n...	...p(a|v)×p(迅速|a)×p(n|a)×p(报告|n)×p(n|n)...	1.20695769e-9
...v a v n...	...p(a|v)×p(迅速|a)×p(v|a)×p(报告|v)×p(n|v)...	2.8957414e-11
...v a vn n...	...p(a|v)×p(迅速|a)×p(vn|a)×p(报告|vn)×p(n|vn)...	1.0085986e-10
...v an n n...	...p(an|v)×p(迅速|an)×p(n|an)×p(报告|n)×p(n|n)...	8.2437876e-12
...v an v n...	...p(an|v)×p(迅速|an)×p(v|an)×p(报告|v)×p(n|v)...	2.4765193e-12
...v an vn n...	...p(an|v)×p(迅速|an)×p(vn|an)×p(报告|vn)×p(n|vn)...	3.0403464e-12

... 当事人/n 应当/v 迅速/ad 报告/v 公安/n 机关/n ...

改进基于HMM的词类标注

- 暗含两个假设：

  (1)句中某个词是否出现只和该词的词类标记有关。和句中的其他词以及其它词的词类标记无关。

  (2)句中某个词的词类只和该词前面一个词的词类有关。而和句中其它词类无关。(词类的bigram模型)

- 可以扩充基于隐马尔科夫模型的词类标注模型，考虑更多的上下文，把词类的bigram模型改作trigram模型。

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}}\, P(t_1^n \mid w_1^n) = \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} p(w_i \mid t_i) p(t_i \mid t_{i-1}, t_{i-2})$$

改进基于HMM的词类标注

- 最大似然估计

$$p(t_i \mid t_{i-1}, t_{i-2}) = \frac{c(t_{i-2}, t_{i-1}, t_i)}{c(t_{i-2}, t_{i-1})}$$

- 数据稀疏问题、应用平滑技术(线形插值)

$$\hat{p}(t_i \mid t_{i-1}, t_{i-2}) = \lambda_1 p(t_i \mid t_{i-1}, t_{i-2}) + \lambda_2 p(t_i \mid t_{i-1}) + \lambda_3 p(t_i)$$

- 输出概率平滑

$$p(w \mid t) = \frac{c(t, w) + 1}{c(t) + T_w}$$

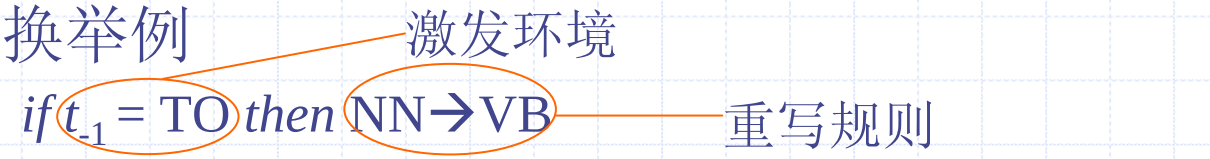
基于转换的词类标注

- Eric Brill提出(1995)
- 特点(兼具规则和统计两个方面的特性)
  - 应用规则进行标注，规则称为转换。
  - 规则不是人工总结，而是应用机器学习的办法学习得到。使用的机器学方法通常称作基于转换的学习(Transformation-Based Learning or TBL)。

☞ Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, Computational Linguistics, Vol. 21, No. 4, 1995, pp.543-565

基于转换的词类标注

- 什么是一个转换(transformation)？
  - 激发环境(triggering environment)
    - 描述了应用该转换需要满足的条件
  - 重写规则(rewriting rule)
    - 描述了应用规则所要进行的动作
    - 重写规则形如 $t_1 \rightarrow t_2$，含义是把词类标记t_1改作t_2
    - 注意重写规则与一般意义上的重写规则的区别
- 转换举例



含义：*Change* NN *to* VB *when the previous tag is* TO

基于转换的词类标注

- 转换的应用
  - *race* NN VB
    - … is expected to **race** tomorrow …
    - …the **race** for outer space …
  - 初标注结果
    - … is/VBZ expected/VBN to/TO race/NN tomorrow/NN …
    - … the/DT race/NN for/IN outer/JJ space/NN …
  - 应用转换规则
    - … is/VBZ expected/VBN to/TO race/VB tomorrow/NN …
    - … the/DT race/NN for/IN outer/JJ space/NN …
  - 转换规则可以视为一种纠错规则
    - 在转换规则使用前，待标注的句子已经进行过初步标注，转换规则负责改正其中的错误标注

基于转换的词类标注

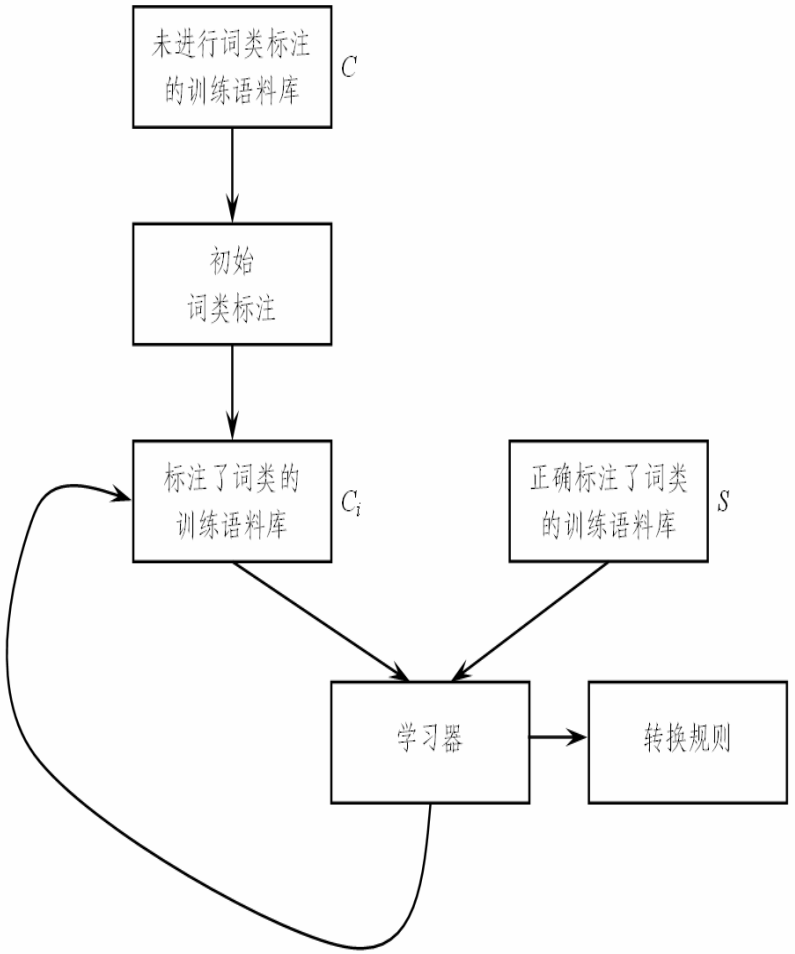
激发环境： 当前词前面一个词的词类是副形词(ad)
重写规则： 把当前词的词类从名词(n)改作动词(v)

… 当事人/n 应当/v 迅速/ad **报告**/n 公安/n 机关/n …
… 当事人/n 应当/v 迅速/ad 报告/v 公安/n 机关/n …

… 发动/v 全区/n 企事业/n 单位/n 积极/ad 文化/n 扶贫/vn …
… 发动/v 全区/n 企事业/n 单位/n 积极/ad 文化/v 扶贫/vn …

基于转换的词类标注

- 学习转换的基本思想
  - 准备未加标注的训练语料
  - 对训练语料进行初标注，形成语料C_0
  - 对C_0进行人工校对，形成正确标注的语料库C
  - 将C_0与C进行对比，学习转换规则
  - 评价学到的转换规则，选择能最大限度地降低C_0错误率的规则τ
  - 对C_0应用转换τ，产生语料C_1
  - 对比C_1与C，按照上述过程继续学习、应用转换规则，直到错误率不再有明显降低为止



基于转换的词类标注

PROCEDURE *TBLearner*(S, T) **begin**

$C \leftarrow$ 删除 S 中的词类信息形成的未标注词类的语料库；

$C_0 \leftarrow$ 基于初始标注程序对 C 进行标注形成的标注语料库；

for $k \leftarrow 0$ **step** 1 **do**

$\tau \leftarrow$ 可使 $E(u_i(C_k))$取最小值的转换 u_i；

if ($E(C_k) - E(\tau(C_k)) < \varepsilon$) **then break**

$C_{k+1} \leftarrow \tau(C_k)$;

$T_{k+1} \leftarrow \tau$;

end

end.

基于转换的词类标注

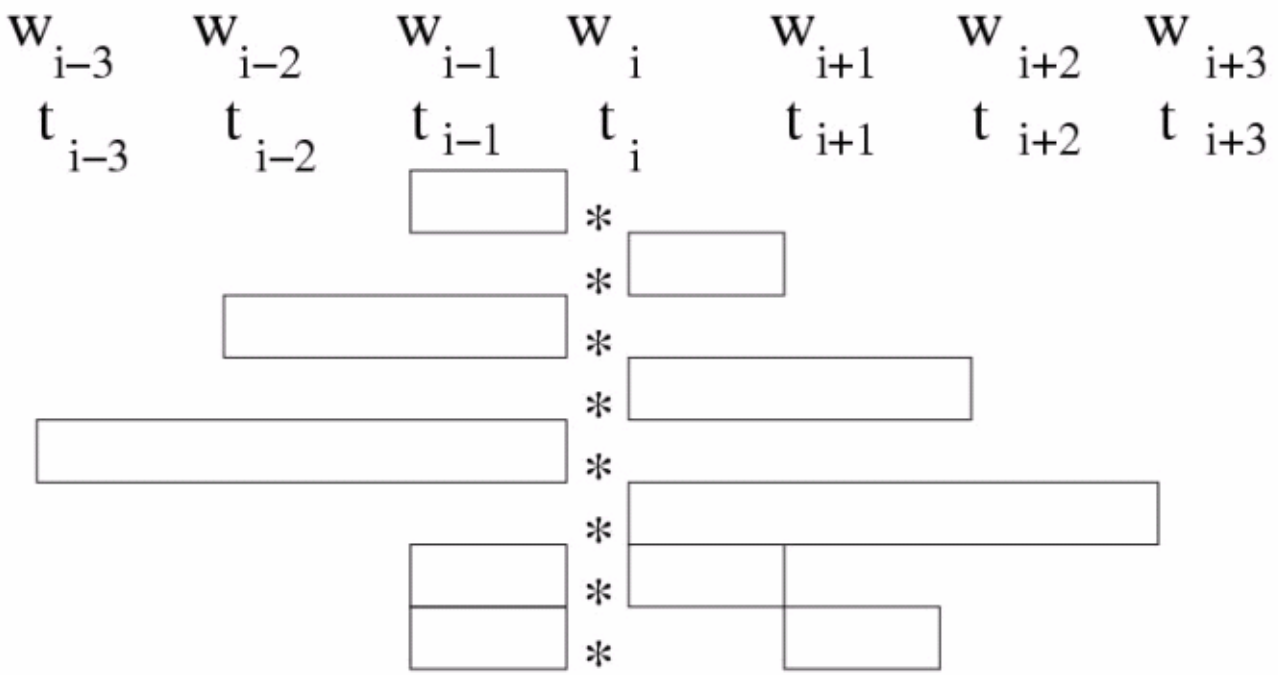
- 初标注器的选择
  - 学习到的转换规则和初标注器有关，选择不同的初标注器学习到的转换规则不同
  - 用dumb tagger进行初始标注
    用基于规则的词类标注器进行标注
    用基于隐马模型的词类标注器进行标注
    …
  - 用学到的规则进行词类标注时，应保证和学习规则适用相同初标注器。
- 转换规则的排列顺序是有意义的
  - 先学到的转换规则先使用，后学到的规则后使用，后学到的规则的作用对象是先学到的规则的处理结果
  - 先学到的规则效果明显、后学到的规则对错误率的改进较小
  - 规则的使用过程类似于创作油画

基于转换的词类标注

- 激发环境的选择
  - 激发环境的选择确定了利用的上下文知识的多少
    - 前文例子中，激发环境仅考虑了标注词前一个词的词类信息(即 t_{-1} = TO)
  - 理论上，利用的上下文知识越多性能越好
  - 对激发环境不加限制，导致学习效率严重下降，需进行权衡
  - Brill使用激发环境模板来限制可以使用的环境

	Change tags			
#	From	To	Condition	Example
1	NN	VB	Previous tag is TO	to/TO race/NN → VB
2	VBP	VB	One of the previous 3 tags is MD	might/MD vanish/VBP → VB
3	NN	VB	One of the previous 2 tags is MD	might/MD not reply/NN → VB
4	VB	NN	One of the previous 2 tags is DT	
5	VBD	VBN	One of the previous 3 tags is VBZ	

基于转换的词类标注



Brill Tagger 中使用的激发环境模板

未登录词

- 未登录词
  - 视作兼类词，可能是任何一个词类，均匀分布
  - 依照出现一次的词(hapax legomenon)的规律处理
    - 更可能是名词 不大可能是限定词等
    - 将出现一次的词的分布平均作为未登录词的分布
  - 对于英文等语言可以利用形态特性(词缀)、拼写特性判定(首字母大小写)

其他方法

- 在上世纪90年代，词类标注问题得到了持续关注，不断有新的方法和模型提出，除我们介绍的方法外，下列方法也取得了较好的标注效果：
  - 基于决策树(Schmid 1994)
  - 基于神经元网络(Benello et al. 1989)
  - 基于最大熵原则(Ratnaparkhi 1996)

[1] Schmid,H., Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. 1994.

[2] Benello, J., et al. Syntactic category disambiguation with neural networks. Computer Speech and Language, 3, 1989.

[3] Ratnaparkhi, A., A Maximum Entropy Part of Speech Tagger. In conference of Empirical Methods in Natural Language Processing, University of Pennsylvania, 1996.