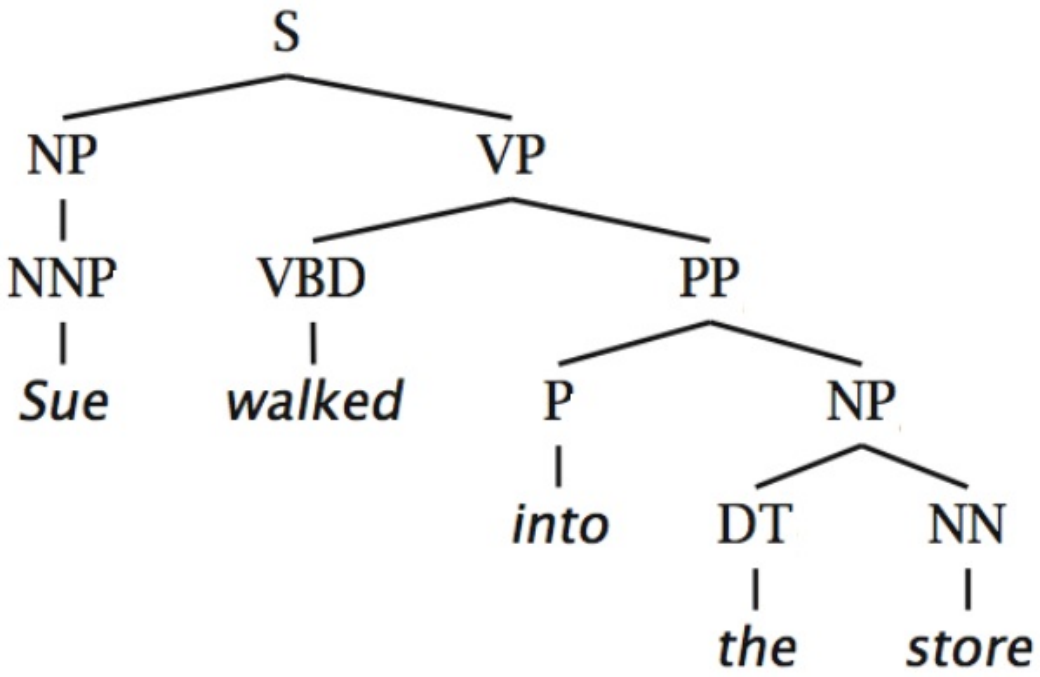
斯坦福公开课学习笔记Part 14-18

NLP

14. 词汇化的PCFGs

14.1 介绍

PCFG的一个问题是他只包含了词性，而没有包括词汇的内容。例如，我们已知VP->VBD PP的概率为0.3，但这只是一个很泛的概率，实际上当动词不同时，所对应的VP->VBD PP概率一定是不一样的。例如，see后面接介词短语的可能性很小，但walk接介词短语的可能性就非常大。
为了解决这个问题，词汇化的PCFG不仅考虑到了词性，还将引入了中心词(head word)，VP[walk]->VBD PP和VP[see]->VBD PP的概率将是不一样的。



14.2 Charniak模型

Charniak模型在计算条件概率时是自上而下的，实际语法分析过程与CKY很相似是自下而上的。
Charniak模型是在自上而下地不断寻找中心词和规则。
计算中心词：P(h|ph,c,pc)，其中h代表中心词，c代表目前节点的类别，ph代表父节点中心词，pc代表父节点类别
计算规则：P(r|h,c,pc)，在找到中心词后，利用中心词来计算规则r

不同的中心词概率不同

单词汇概率：当中心词概率不同时，规则的转换概率也是不同的

Local Tree	*come*	*take*	*think*	*want*
VP → V	9.5%	2.6%	4.6%	5.7%
VP → V NP	1.1%	32.1%	0.2%	13.9%
VP → V PP	34.5%	3.1%	7.1%	0.3%
VP → V SBAR	6.6%	0.3%	73.0%	0.2%
VP → V S	2.2%	1.3%	4.8%	70.8%
VP → V NP S	0.1%	5.7%	0.0%	0.3%
VP → V PRT NP	0.3%	5.8%	0.0%	0.0%
VP → V PRT PP	6.1%	1.5%	0.2%	0.0%

双词汇概率：也可以利用其它特征来计算单词的条件概率，比如在WSJ中，P(prices)出现的概率可能不高，不过当我们已知单词为名词复数形式时，P(prices|n-plural)的概率就是1.3%；而当我们已知单词为名词短语中心词、后面紧接动词过去时、后面紧接fell等信息后，P(prices|conditions)的概率就变成了14.6%

- P(prices | n-plural) = .013
- P(prices | n-plural, NP) = .013
- P(prices | n-plural, NP, S) = .025
- P(prices | n-plural, NP, S, v-past) = .052
- P(prices | n-plural, NP, S, v-past, fell) = .146

线性插值法

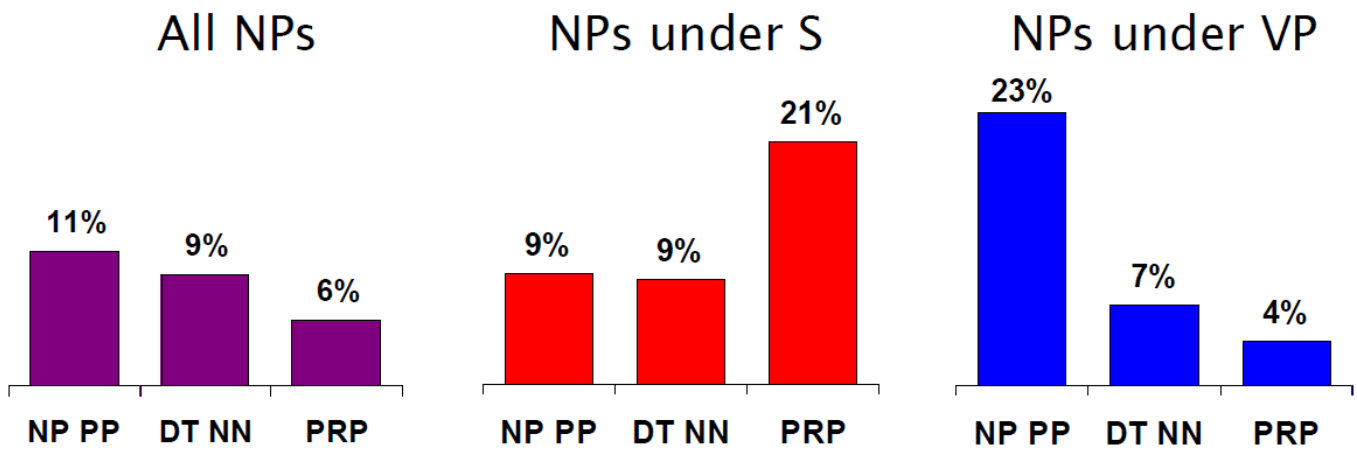
实际上也是一种平滑方法，在训练样本不够多时十分有效。为了防止P(h|ph,c,pc)为0，可以将概率估算值记为

$$\hat{P}(h|ph,c,pc) = \lambda_1(e)P(h|ph,c,pc) + \lambda_2(e)P(h|C(ph),c,pc) + \lambda_3(e)P(h|c,pc) + \lambda_1(e)P(h|c)$$

14.3 独立假设

PCFG有一个很强的独立性假设：在任意节点，已知该节点的类型之后，该节点子树内部的概率与子树外部的概率相互独立，也就是说子节点继续分裂的过程中，概率不受其父节点的影响

这种假设在实际中是存在问题的，因为子节点的分裂方式其实受父节点的影响很强



解决这个问题的一个方法，是在标注类别的时候也同时加上单词父节点的类别，不过这也会导致特征更加稀疏

14.4 非词汇化PCFG

非词汇化PCFG不会使用词汇方面的特征，而是只考虑语法方面

- 例如，NP-stock不算非词汇化PCFG，因为包括了stocks这种具体的词汇；NP^S-CC就属于非词汇化PCFG
- 一些语言学中的常用单词是被允许使用的，比如VB-have等