结构上下文相关的概率句法分析*

张浩[1] 刘群[1] 白硕[1,2]

[1](中国科学院 计算技术研究所,北京 100080)

[2](国家计算机与网络信息安全管理中心,北京 100031);

E-mail: zhanghao@software.ict.ac.cn

摘要: 本文研究了 PCFG 独立性假设的局限性，并在 PCFG 的基础上提出了三个逐层递进的与结构上下文相关的概率句法分析模型，它们考虑了分析树当中每个派生节点的结构上下文条件。为了更好地说明方法本身的问题，我们在宾州中文树库和一个短句树库上都进行了实验，文中给出了横向和纵向的对比实验数据。结果表明，系统地将结构上下文条件引入的做法以很小的代价提高了概率句法分析器的性能，值得推广和深入研究。

关键词: 概率句法分析; 结构上下文; Chart 分析; 汉语自动分析

引言

自然语言处理当中的句法分析是以建立起一个句子的树型结构为目标的分析过程。建立起这样的树结构要求首先建立起一套语法规则。最早的规则系统就是简单的上下文无关语法规则。用这样的规则去分析自然语言的句子，歧义问题难以避免。一个句子常常会被分析出成千上万的符合规则的分析树来。为了把规则描述得更加细致，各种由上下文无关语法派生出来的语法规则系统又被引入这个领域。概率上下文无关语法(PCFG)就是把概率引入到上下文无关语法规则系统而形成的语法规则系统。概率可以为分析过程中的歧义消解提供依据。现在，我们可以选择出一个概率最大的分析树作为分析结果了。但是，经典的 PCFG 实际上是建立在一些非常理想化的独立性假设的基础之上的[1]，而这些假设并不符合实际，于是造成了 PCFG 的实际效果不理想。本文的工作对树库当中分析树的局部结构进行了统计，估计出了在一定周边结构条件限制下的规则施用概率。这在一定程度上是从语法结构的层面对 PCFG 的上下文无关假设所进行的突破。

本文第一部分将介绍基于上述思想的几个层次的概率模型，并说明它们之间的关系以及它们与经典 PCFG 的关系。第二部分将介绍实验的具体步骤、所用算法。第三部分将针对实验结果进行分析说明。

1 概率模型

1.1 经典概率模型——PCFG

一个 PCFG G 的符号系统包括以下成分：

- 一个终结符的集合，$\{w^k\}$, $k = 1, ..., V$

* 本课题受国家重点基础研究项目（973）资助（G1998030510 和 G1998030507-4）

- 一个非终结符的集合，$\{N^i\}, i=1,...,n$
- 一个开始符号，N^1
- 一个规则的集合，$\{N^i \to \zeta^j\}$，（其中ζ^j是一个终结符和非终结符的序列）

概率方面，PCFG 给出了由一个非终结符节点可能派生出的各种符号序列的概率分布，即：

$$\forall i \quad \sum_j P(N^i \to \zeta^j \mid N^i) = 1$$

计算一棵分析树t的概率$P(t)$，需要进行必要的独立性假设。经典的 PCFG 认为，施用每一条规则的概率独立于上下文和祖先节点[1]。或者说，给定了一个非终结符节点，它会以多大概率派生出什么样的子节点序列，是语法决定的。这样，假设t当中的所有规则构成了一个规则的多重集R，那么：

$$P(t) = \prod_{r \in R} P(r \mid LHS(r))$$

1.2 考虑祖先节点条件的概率模型——P-PCFG

符号系统不变，改变独立性假设，认为每条规则的施用概率要受到左部非终结符节点的父节点的决定。相应地，模型中包括了这样的概率分布：

$\forall i,k \sum_j P(N^i \to \zeta^j \mid < N^i, N^k >) = 1$，其中，$< N^i, N^k >$属于$N^i$是$N^k$之子的关系。

分析树的概率计算公式：

$$P(t) = \prod_{r \in R} P(r \mid < LHS(r), ParentOf(LHS(r)) >)$$

短语的结构是受到上层节点的制约的。例如，做主语的 *NP* 短语（*NP* 位于 *S* 之下）和做宾语的 *NP* 短语（*NP* 位于 *VP* 之下）的内部结构有着明显不同的概率分布[1]。P-PCFG 把这一决定因素加入到了概率模型当中。

1.3 考虑祖先节点和节点位置条件的概率模型——PORD-PCFG

认为每条规则的施用概率要受到左部非终结符节点的父节点和左部非终结符节点在兄弟节点中的排行来决定。相应地，模型中包括了这样的概率分布：

$\forall i,k,ord \quad \sum_j P(N^i \to \zeta^j \mid < N^i, N^k, ord >) = 1$，其中，$< N^i, N^k, ord >$属于$N^i$是$N^k$的第 ord 子的关系。

分析树的概率计算公式：

$$P(t) = \prod_{r \in R} P(r \mid < LHS(r), ParentOf(LHS(r)), OrderOf(LHS(r)) >)$$

如果一个非终结符节点下面派生出了两个或两个以上相同的非终结符节点，它们唯一的区别是位置不同，位置信息这时候就非常关键了。例如，如果一个动词短语 *VP* 带两个宾语 $NP-1$ 和 $NP-2$，那么这两个宾语的结构将有着明显的差异[1]。PORD-PCFG 同时顾及到了父节点和排行位置对短语结构的影响。

1.4 考虑上一级规则的概率模型——PRORD-PCFG

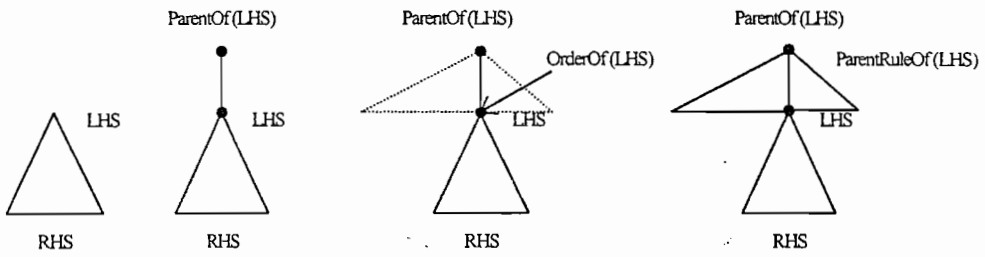
认为每条规则的施用概率要受到左部非终结符节点所在施用规则和在施用规则中所处的位置来决定。相应地，模型中包括了这样的概率分布：

$\forall i,r,ord \ \sum_j P(N^i \to \zeta^j \mid < N^i, r, ord >) = 1$，其中，$< N^i, r, ord >$ 属于规则 r 右部第 ord 子为 N^i 的关系。

分析树的概率计算公式：

$$P(t) = \prod_{r \in R} P(r \mid < LHS(r), ParentRuleOf(LHS(r)), OrderOf(LHS(r)) >)$$

PRORD-PCFG 完全考虑了派生链中的上一级的规则对本级规则的干预。上一级的规则给出了一个相对封闭的局部上下文环境，父节点、兄弟、排行位置的作用都包括在其中。



(1) $P(r \mid LHS(r))$

(2) $P(r \mid < LHS(r), ParentOf(LHS(r)) >)$

(3) $P(r \mid < LHS(r), ParentOf(LHS(r)), OrderOf(LHS(r)) >)$

(4) $P(r \mid < LHS(r), ParentRuleOf(LHS(r)), OrderOf(LHS(r)) >)$

图 1 给出了四个模型的示意，表示了结构上下文是如何被渐进式地加入到概率模型当中的。

2 实验步骤和算法

这一部分将介绍语法规则及其概率的获取，以及分析算法和结果评测算法。

2.1 语法规则及相应概率的获取

从树库自动抽取语法规则并进行统计以构造 PCFG 的方法在英语的句法分析领域已经被证明是简单易行的[2]。中文树库资源目前还非常稀少，这也是汉语概率句法分析方面的研究开展不多的一个重要原因。宾州中文树库[3]目前已经初具规模，句子都来自新华社的新闻稿，句长基本都在 30-100 词之间。虽然其规模还比较小，但是毕竟这是一个公共的资源，其中的句子具有真实性，应该可以作为各种方法的试金石。我们自己也有一个短句树库[4]，来源是我们的一个机器翻译系统的正确分析结果，其中的句子基本涵盖了常见语法现象，但是句长大约都在 5-20 个词之间，缺乏真实性。我们在两个树库上都做了实验，以利于说明方法的问题。

PCFG 的提取是很简单的。首先统计训练语料当中出现的规则及其出现次数。然后用最大似然估计从规则出现频率估计出规则施用概率[2]：

$$\hat{P}(N^i \to \zeta^j) = \frac{C(N^i \to \zeta^j)}{\sum_k C(N^i \to \zeta^k)} \quad (1)$$

加入结构上下文条件的规则条件概率可以通过对树库当中的节点标记进行变换从而归结到不考虑上下文条件的 PCFG 的规则概率。标记变换方法如下：

(1) 对于 P-PCFG，将树库中非根且非叶的中间节点符号尾部加上其父节点符号作为后缀。

(2) 对于 PORD-PCFG，将中间节点符号加上父节点符号和排行位置作为后缀。

(3) 对于 PRORD-PCFG，加上上一级的规则和排行位置作为后缀。

我们用的方法和[5]的方法是一致的，但比之更为通用化了。

示例如下图：

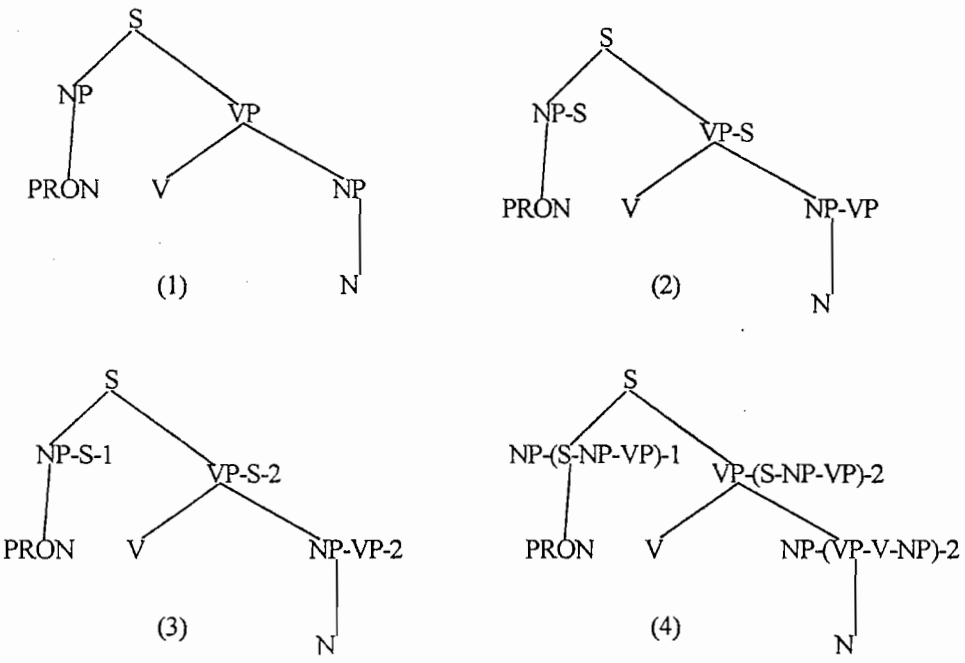


图 2 标记变换示例

完成了标签变换之后，结构上下文条件就都由后缀来表示了。我们仍使用前面的规则抽取方法和统计方法，将得到另外 3 组规则。细化的条件概率就是细化后的规则所对应的概率。

这样，我们的分析算法都是 PCFG 的分析算法，只要替换不同的规则集，就构造出了与一定局部上下文相关的分析器。在分析结束之后，很容易将标签变换回来，只要去掉后缀。

2.2 分析算法和结果评测算法

分析算法就是找到一个 Viterbi 分析的过程[1]。我们的分析算法是在一个优化过的 Earley 算法[6]的基础之上增加了子树概率最大化运算的一个算法。分析算法是自左向右进行的，每当一个子树生成，就要进行局部最优的计算，保留概率最大的歧义派生。最后得到的完整的分析树就是由局部最优派生构成的全局最优派生。

结果评测算法我们采用了PARSEVAL[7]，计算标签召回率(LR)、标签精确率(LP)、平均交叉括号数(CBs)，0-括号交叉率(0CB)，1-括号交叉率(1CB)。具体的计算公式这里不再列出。

3 实验结果

本节将给出几个模型的实验结果并加以分析。

表 2-1 机器翻译系统树库封闭（1-2400 句）和开放（2401-3072 句）测试结果

	PCFG		P-PCFG		PORD-PCFG		PRORD-PCFG	
	封闭	开放	封闭	开放	封闭	开放	封闭	开放
LP	87.56	85.58	90.49	87.60	91.90	88.47	94.03	89.30
LR	87.44	85.66	91.31	88.53	93.66	90.83	94.71	91.25
CBs	0.71	0.86	0.51	0.63	0.39	0.50	0.30	0.46
0CB	64.00	58.57	74.46	69.02	79.92	74.66	85.25	75.85
1CB	80.25	76.30	85.67	82.56	88.83	87.33	91.54	87.28

表 2-2 宾州中文树库封闭（1-2200 句）和开放（2201-2863 句）测试结果

	PCFG		P-PCFG		PORD-PCFG		PRORD-PCFG	
	封闭	开放	封闭	开放	封闭	开放	封闭	开放
LP	75.65	75.71	79.31	77.76	81.55	77.73	87.97	74.91
LR	69.99	70.27	73.93	73.02	77.48	75.3	85.83	76.42
CBs	4.18	3.42	3.40	2.90	2.96	2.79	1.72	3.18
0CB	19.90	24.81	25.32	31.16	28.06	31.56	46.91	29.62
1CB	34.32	40.24	41.32	49.47	46.47	48.1	65.68	45.42

上面的结果表明，引入结构上下文信息的确提高了分析器的性能。在基本上都是短句，训练数据相对充足的机器翻译系统树库上面的实验结果尤其令人满意。在宾州中文树库的开放测试中，后两个模型的实验结果中某些指标开始下降。我们分析认为，数据稀疏问题严重影响了结果，而我们的方法还有改进的余地，因为目前的结果是在没有使用任何平滑技术情况下得到的结果。在接下来的工作中，我们准备采用 backoff 的平滑技术，把几个层次的结构上下文信息协调起来加以利用，相信会得到好的结果。

4 总结和展望

PCFG 上下文无关的假设，包括结构上下文无关和词汇上下文无关都是必须加以突破的。我们从相对容易的结构上下文相关性入手，已经有了一定收获。从国际主流的句法分析方法来看，将句法分析器词汇化，把词汇信息渗透到分析的各个层次，才能有更为本质上的性能提升。但是，中文树库的规模限制了对各种方法的尝试。希望中文树库的建设能够尽快跟上步伐。在资源有限的情况下，我们也必须考虑充分利用现有的资源，挖掘其中的价值。

参考文献:

[1] Manning, C., Schütze, H. 1999. Foundations of Statistical Natural Language Processing. MIT Press.
[2] Charniak, E. 1996. Treebank grammars. Technical Report CS-96-02, Department of Computer Science, Brown University.

[3] Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. Developing guidelines and ensuring consistency for Chinese text annotation. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC-2000), Athens, Greece, 2000.
[4] 刘颖. 规则方法和统计方法相结合在汉英机器翻译中的研究和应用. 博士论文. 中科院计算所. 1998.
[5] Johnson, M. 1998. The effect of alternative tree representations on tree bank grammars. In Proceedings of the Joint Conference on New methods in Language Processing and Computational Natural Language Learning (NeMLaP3/CoNLL'98), pp 39—48.
[6] 白硕，张浩. 角色反演算法. 软件学报. 已录用. 2002.
[7] Charniak, E. 1997. Statistical parsing with a context-free grammar and word statistics. In Proceedings of NCAI-1997, pp 598—603.

致谢 感谢计算所软件室自然语言处理组的所有成员，大家的热烈讨论使作者受益非浅。特别感谢李继锋同学和作者一起进行分析器的调试。特别感谢张华平同学提供了许多关于论文写作的意见，特别感谢李素建博士在树库等问题方面给予的帮助。

作者简介： 张浩(1978－)，男，山西孝义人，硕士生，主要研究领域为自然语言处理；刘群（1966－），男，江西萍乡人，在职博士生，副研究员，主要研究领域为机器翻译，自然语言处理与中文信息处理；白硕(1956－)，男，辽宁辽阳人，研究员，博士生导师，主要研究领域为自然语言处理、网络安全

Structural Context Conditioned Probabilistic Parsing of Chinese*

ZHANG Hao[1] LIU Qun[1] BAI Shuo[2]

1(Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)
2(National Administrative Center for Network and Information Security, Beijing 100031, China);
E-mail: zhanghao@software.ict.ac.cn

Abstract: Three probabilistic parsing models, which are successive augmentations of the conventional PCFG, are presented in this paper. In this sequence of models outlined, wider and wider structural context is taken as the conditioning events to condition the derivations. We have applied the models to the task of Chinese parsing. To reveal the problems of the method more objectively, results on the Penn Chinese Treebank and another treebank composed mainly of short sentences are both reported. The results show that the Labeled Precision Rates and Labeled Recalling Rates are raised gradually in this approach. We suggest that with smoothing techniques taken, even better results can be expected.

Key words: Probabilistic Paring; Structural Context; Chart Parsing; Chinese NLP

* Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1998030510 &G1998030507-4.