**读书报告**

王贵涛

自然语言处理的第十二章主要介绍了选区语法（Constituency Grammars）。

**树库（Treebanks）**

由上下文无关的语法规则组成的足够健壮的语法可用于为任何句子分配解析树。树库中的每个句子都与相应的解析树配对。

树库中的句子隐含地构成了被注释的语料库所代表的语言的语法。树库有大量的平面规则，给概率解析算法带来了问题，由于这个原因，通常会对从树库中提取的语法进行各种修改。

句法成分可以与词汇头相关联。在特定句子的树的上下文中动态的识别头部，一旦句子被解析，生成的树将被遍历，以用适当的头装饰每个节点。

**语法等价与范式（Grammar Equivalence and Normal Form）**

一种形式语言被定义为一组字符串。这意味着我们可以通过询问两种语法是否生成相同的字符串集来询问它们是否等价。实际上，可以用两种不同的上下文无关语法生成同一种语言。

语法对等通常分为两种：弱对等和强对等。如果两种语法生成相同的字符串集，并且为每个句子分配相同的短语结构，那么它们是强对等的。如果两种语法生成相同的字符串集，但不为每个句子分配相同的短语结构，则这两种语法是弱对等的。

有时，为语法提供一种标准形式是有用的，在这种形式中，每个结果都采用一种特定的形式。

**词汇化语法（Lexicalized Grammars）**

目前的语法方法强调短语结构规则，而尽量减少词汇的作用。然而，这种方法导致的解决方案最多是麻烦的，产生的语法是冗余的、难以管理的和脆弱的。

为了克服这些问题，人们开发了许多备选的词典，它们都有一个共同的主题，那就是使用词典。与计算相关的方法在词汇化程度、依赖词汇而不是短语结构规则来捕获语言事实的程度方面有所不同。