**浅层句法分析方法综述**

**摘要** ：浅层句法分析是自然语言处理领域出现的一个重要的语言处理策略。它不要求得到完全的句法分析树,只要求识别其中的某些结构相对简单的成分。它将句法分析分解为两个子任务：（1）语块的识别和分析（2）语块之间的依附关系分析。浅层句法分析的首要任务是语块的识别和分析。本文着重介绍英语浅层句法分析中所应用的一些技术，包括基于统计的方法和基于规则的方法。

**关键词** ：浅层句法分析 统计方法 互信息方法 概率方法

**1.引言**

浅层句法分析，也叫部分句法分析或语块分析,是近年来 自然语言处理领域 出现的一个新的语言处理策略。它是跟完全句法分析相对的。传统的句法分析要求通过一系列分析过程，最终得到句子的完整的句法树。而浅层句法分析则不要求得到完全的句法分析树它只要求识别其中的某些结构相对简单的成分如非递归的名词短语、动词短语等。这些识别出来的结构通常被称作语块，语块和短语这两个概念可以换用。浅层句法分析的结果并不是一棵完整的句法树，但各个语块是完整句法树的一个子图,只要加上语块之间的依附关系,就可以构成完整的句法树。所以浅层句法分析将句法分析分解为两个子任务语块的识别和分析语块之间的依附关系分析。浅层句法分析的主要任务是语块的识别和分析。这样就使句法分析的任务在某种程度上得到简化,同时也有利于句法分析技术在大规模真实文本处理系统中迅速得到应用。概括起来,句法分析的方法基本上可以分成两类基于统计的方法和基于规则的方法。当然也可以采用规则和统计相结合的混合方法。下面第2节介绍基于统计的方法,第3节介绍基于规则的方法。

**2. 基于统计的方法**

近年来,随着语料库技术的发展,许多统计方法被用在短语识别和分析上。这些方法主要来自概率统计和信息论,以下介绍其中影响较大的几种方法：（1）基于隐马尔科夫模型的方法（2）互信息方法护（3）统计方法（4）基于中心词依存概率的方法。

**2.1 基于隐马尔科夫模型的方法**

隐马尔科夫模型,是从语音识别中发展出来的一种统计技术,它提供了一种基于训练数据提供的概率来自动构造识别系统的技术。一个隐马尔科夫模型包含两层一个可观察层和一个隐藏层这个隐藏层是一个马尔科夫过程,即是一个有限状态机,其中每个状态转移都带有转移概率。在语音识别中,可观察层是声音片段的序列,隐藏层是用音素序列表示的词的发音的有限状态模型。用口语录音片段及其转写坦作为训练数据训练，它就可以用作识别器

用于识别未训练过的声音片段,从而生成口语的转写形式。计算语言学家最早把 技术应用于英语的词性标注,并取得了极大的成功,仅依靠简单的统计方法就可以达到左右的正确率。在词性标注中,可观察层是词的序列,隐藏层是词类标记的序列,训练数据是标注好词性的文本语料,经过训练的就成为自动标注器,它可 以给只包含词序列的文本中的每个词标注上词类标记。

**2.2 互信息方法**

互信息是信息论中的一个概念,它用来度量一个消息中两个信号之间的相互依赖程度。我们可以把词类序列看成随机事件，这样就可以计算一对词类标记之间的互信息。为了确定句子中短语的边界，不能局限于两个符号的组合内部的互信息,需要看更多的上下文,即把二元互信息扩展为一个符号的组合内部的互信息。提出了广义互信息的概念,它根据两个相邻的词类标记的上下文在一个观察窗口内来决定它们之间是否是一个短语边界所在。在下面的公式推导中,表示二元互信息。

**2.3 统计方法**

用统计方法来度量两个词之间的关联度。用这种方法来确定短语的边界。

**2.4 基于中心词依存概率的方法**

提出了一种基于分析树中中心词之间依存概率的统计分析算法,该方法的要点是,把分析树归结为一个非递归的基本名词短语,简称集合及依存关系的集合。

所以依存关系就是中心词和其他词之间的依存关系。依存概率可以通过树库中的统计得到。分析算法是一个自底向上的线图分析器,利用动态规划来查找训练数据中所有的依存关系空间。例如,由句子的分析树可以得到的集合及中心词之间的依存关系集合。

**3 基于规则的方法**

规则方法就是根据人工书写的或 半 自动获取的语法规则标注出短语的边界和短语的类型。根据标注策略的不同可以把规则方法分为两种增加句法标记法,即在词串中插入短语边界和短语类型等句法标记删除句法标记法,即从多个候选的句法标记中删除不合法的标记。

**3.1 增加句法标记法**

增加句法标记的句法分析包括一个状态转换器序列,转换器由正则式构成,即语法规则是有限状态语法的形式。大部分的规则系统都采用这种方法,如的语块分析系统。有限状态层叠包括多个层级,分析逐层进行。每一级上短语的建立都只能在前一级的基础之上进行,没有递归,即任何一个短语都不包含同一级的短语或高一级的短语。分析过程包括一系列状态转换,用表示。通常的状态转换操作的结果是在词串中插入句法标记,而有限状态层叠则在每一级转换上用单个的元素来替换输入串中的一个元素序列,就跟传统的句法分析一样。每一个转换

定义为一个模式的集合。每一个模式包括一个范畴符号和一个正则式。正则式转换为有限状态自动机模式。自动机合在一起就产生一个单一的、确定性的有限状态层级识别器一它以级的输出为输入,并产生作为输出。在模式匹配过程中,如遇到冲突即两个或两个以上的模式都可以运用,则按最长匹配原则选择合适的模式。

如果输入中的一个元素找不到相应的匹配模式,则把它直接输出,继续下一个元素的匹配。

**3.2 删除句法标记法**

这种方法的思想来自词性标注。在词性标注中,首先从词典中查出每个词可能具有的所有词性,然后根据上下文来消歧,从中选择一个正确的词性。这种思想用到句法标注上就是首先标注出每个词可能的句法功能,然后根据上下文来消歧,从中选择出一个正确的句法功能标记。也就是说,句法分析包括两个主要步骤给出输入词可能的句法功能标记与上下文无关,可能有多个候选删去在上下文中不可接受的句法标记,或从几个候选中选出一个最合理的句法标记即同时排除其他标记。这样,句法分析实际上成了一个删除在上下文中不合法的句法标记的过程。

下面举例说明。

**3.3 语法规则的自动学习**

在基于规则的方法中,主要的困难在于语法规则的获取 以及语法规则之间的优先顺序排列。提出了一种基于转换的错误驱动的学习方法,这种方法首先被用于词性标注,得到的结果可以和统计方法相媲美。把这种自学习方法用于识别英语中的基本名词短语。这种方法通过学习得到一组有序的识别基本名词短语的规则。

另一组语法规则自动获取的方法是采用机器学习中基于实例的方法一或基于记忆的方法仃一,如和电。下面首先介绍基于转换的学习方法,然后介绍基于实例的方法。基于转换的规则学习方法如图所示,基于转换的学习方法以下列三部分资源为基础带标注的训练语料库。对于识别任务来说,训练语料要标注出其中所有正确的在此之前当然要先标注词性。规则模板集合。规则模板集合用于确定可能的转换规则空间。一个初始标注程序。基于转换的错误驱动的学习算法是初始标注。把训练语料中所有的标记去掉,用一个简单的初始标注程序标注出训练集中可能的。把这个结果作为系统的底线。生成候选规则集。在每个初始标注错误的地方,规则模板便用来生成候选规则,规则的条件就是词的上下文环境,动作就是改正错误标记所要做的动作。规则的条件就是词的上下文环境,动作就是改正错误标记所要做的规则的条件就是词的上下文环境,动作就是改正错误标记所要做的动作。

**参考文献 ：**

**俞士汶，2000，浅层句法分析方法概述，当代语言学第2卷2000年第2期；**

**Steven Abney , 1996, Partial Parsing via Finite-State Cascades,**

**In *Proceedings of the ESSLLI’96 Robust Parsing Workshop.***