《当代语言学》第2卷2000年第2期74-83页，北京

浅层句法分析方法概述

孙宏林　俞士汶　　北京大学计算语言学研究所

提要 浅层句法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。浅层句法分析的主要任务是语块的识别和分析。90年代以来，国外在英语的浅层句法方面做了不少工作，国内也有一些学者采用英语中的方法探索汉语的浅层句法分析问题。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(shallow parsing)，也叫部分句法分析(partial parsing)或语块分析(chunk parsing)，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(chunk)，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(subgraph)，只要加上语块之间的依附关系(attachment)，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：（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。

90年代以来，国外在英语的浅层句法分析方面做了不少工作，国内也有一些学者采用英语中的方法探索汉语的浅层句法分析。本文主要就在英语浅层句法分析中所应用的一些技术进行简要的介绍，并简单介绍汉语的相关研究。其中有些方法虽然是面向完全句法分析的，但由于其对完全句法分析的任务进行了分解，所以其技术也可以归入浅层分析的范畴。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，第4节简要介绍汉语的有关研究，最后是结束语。

2. 基于统计的方法

近10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；（3）ϕ^2 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(HMM)的方法

隐马尔科夫模型(Hidden Markov Models, HMMs)是从语音识别中发展出来的一种统计技术(Rabiner 1989)，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(transcription)作为训练数据训练HMM，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HMM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95%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HMM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。

Church(1988)进一步把HMM用于识别英语中简单的非递归的名词短语，他把短语边界识别化为一个在词类标记对之间插入NP的左边界（"["）和NP的右边界("]")的问题。如果不考虑空短语（即"[]"）和短语的嵌套（如"[[", "]]", "][" 等），那么在一对词类标记之间只有四种情况：(1)[；(2)] ；(3)][；(4)空（即无NP边界）。最后一种还可以进一步分为两种情况：(a)无NP边界但在NP之内（I)；(b)无NP边界但在NP之外（O)。这样任意一对词类标记之间就存在5种可能的状态：(1)[；(2)] ；(3)][；(4)I ；(5)O。Church的方法是，首先在有词性标注的语料中人工或半自动标注NP边界，以此作为训练数据，然后统计出任意一对词类标记之间出现以上5种状态的概率。统计得到的概率就成为短语边界标注的根据。这实际上把短语边界的识别变成了一个与词性标注类似的问题。如：

输入：	$	the	procecutor	said	in	closing	that	（词序列）
		DT	NN	VB	IN	NN	CS	（词性序列）
输出：	⟨$,DT⟩	⟨DT,NN⟩	⟨NN,VB⟩	⟨VB,IN⟩	⟨IN,NN⟩	⟨NN,CS⟩	（词性标记对）
	|	|	|	|	|	|	
	[	I	]	O	[	]	（NP边界）

2.2 互信息方法

互信息(mutual information)是信息论中的一个概念(Fano 1961)，它用来度量一个消息中两个信号之间的相互依赖程度。二元互信息是两个事件的概率的函数：

$$(1) \qquad MI(X,Y) = \log_2 \frac{P(X,Y)}{P(X) \times P(Y)}$$

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。如果X和Y在一起出现的机会多于它们随机出现的机会，则$P(X,Y) >> P(X) \times P(Y)$，即$MI(X,Y) >> 0$；如果X和Y是随机分布的，则$P(X,Y) \approx P(X) \times P(Y)$，即$MI(X,Y) \approx 0$；如果X和Y是互补分布的，则$P(X,Y) << P(X) \times P(Y)$，即$MI(X,Y) << 0$。互信息值越高，X和Y组成短语的可能性越大，互信息值越低，X和Y之间存在短语边界的可能性越大。

为了确定句子中短语的边界，不能局限于bigram（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n-gram（n个符号的组合）内部的互信息。Magerman and Marcus (1990)提出了广义互信息(generalized mutual information)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，MI表示二元互信息，MI_n是一个向量，它表示n-gram $(x_1 \cdots x_n)$中任意两个部分之间的互信息，表示这个向量中的第k个分量（$1 \leqslant k < n$），它表示$x_1 \cdots x_k$和$x_{k+1} \cdots x_n$之间的二元互信息。

MI_n^k一个n-gram$(x_1 \cdots x_n)$内部有n-1个二分切分点，每一个切分点的二元互信息为：

(2) $$MI_n^k(x_1 \cdots x_n) = MI(x_1 \cdots x_k, x_{k+1} \cdots x_n)$$

(3) $$= \log \frac{p(x_1 \wedge x_n)}{p(x_1 \wedge x_k)\, p(x_{k+1} \wedge x_n)}$$

在公式(3)中，对于每个MI_n^k(k=1, 2, ⋯, n-1)，分子都相同，当分母最大时，互信息值最小。

基于互信息的短语边界划分的理论基础是，在n-gram中，局部广义互信息值最小的一对标记之间就是短语边界所在的位置。理论推导参见Magerman and Marcus(1990)。

在n-gram$(x_1 \cdots x_i, y_1 \cdots y_j)$ $(1 \leqslant i < n,\ 1 < j < n,\ i+j=n)$ 内部，以两个相邻的词类标记x_i和y_1之间为界，共有$i \times j$个二元组合(bigram)，要计算其间的互信息，应当综合考虑每一个bigram之间的二元互信息，因此产生了广义互信息的概念。广义互信息的计算公式是：

(4) $$GMI_{n(i+j)}(x_1 \wedge x_i, y_1 \wedge y_j) = \sum_{\substack{X\text{以}x_i\text{结束} \\ Y\text{以}y_1\text{开始}}} \frac{1}{\sigma_{XY}} MI(X,Y)$$

这里，$GMI_{n(i+j)}(x_1 \cdots x_i, y_1 \cdots y_j)$表示在一个n-gram中两个相邻的元素$x_i$和$y_1$之间的广义互信息，X表示n-gram中以$x_i$结束的词类标记串，Y表示n-gram中以$y_1$开始的词类标记串。σXY是XY中$M_{|XY|}^k$ ($|XY| = i+j$，$0 < k < i+j$)的标准差。

2.3 ϕ^2统计方法

Gale and Church(1991)用ϕ^2统计方法来度量两个词之间的关联度。Chen and Lee(1995)用这种方法来确定短语的边界。

对于两个词w_1和w_2，可以建立如下的联立表(contingency table)：

	W_2	$!W_2$	Σ
W_1	a	b	a+b
$!W_1$	c	d	c+d
Σ	a+c	b+d	a+b+c+d

在上表中，a表示串w_1 w_2出现的次数，b表示不在w_1 w_2中的w_1的出现次数，c表示不在w_1w_2中的w_2的出现次数，d表示既不是w_1又不是w_2的词的次数。a+b是w_1的出现次数，a+c是w_2的出现次数，c+d是非w_1的总词次，b+d是非w_2的总词次，N=a+b+c+d表示语料库中的总词次。根据上面的联立表，ϕ^2统计量定义如下：

$$\phi^2 = \frac{(a \times d - b \times c)^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)} \quad (5)$$

当a=0时，ϕ^2近于0，即当w_1和w_2从不共现时，ϕ^2取极小值。当b=c=0时，ϕ^2=1，即当w_1和w_2总是共现时，ϕ^2取极大值。ϕ^2值越大，说明w_1和w_2共现的机会越多，相反，ϕ^2值越小，则说明w_1和w_2共现的机会越少。

如果把上面的两个词换成两个词类标记，则可以进行标记对之间的ϕ^2统计。进一步推广则可以在一个词类序列的两个子序列之间进行ϕ^2统计。原理的推导与互信息方法相似，此处从略。

2.4 基于中心词依存概率的方法

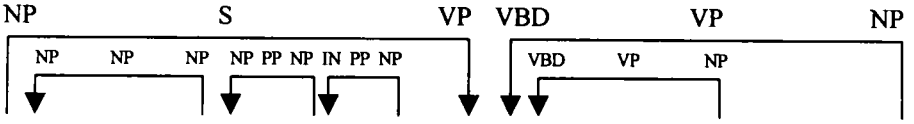
Collins(1996)提出了一种基于分析树中中心词之间依存概率的统计分析算法，该方法的要点是，把分析树归结为一个非递归的基本名词短语(base noun phrase, 简称base NP)集合及依存关系的集合。在这些依存关系中，base NP中除了中心词外其他词都被忽略，所以依存关系就是base NP的中心词和其他词之间的依存关系，依存概率可以通过树库中的统计得到。分析算法是一个自底向上的线图分析器，利用动态规划来查找训练数据中所有的依存关系空间。

例如，由句子(a)的分析树(b)可以得到base NP的集合B及中心词之间的依存关系集合D(c)。

(a) John/NNP Smith/NNP, the/DT president/NN of/IN IBM/NNP, announced/VBD his/PRP $ resignation/NN yesterday/NN.

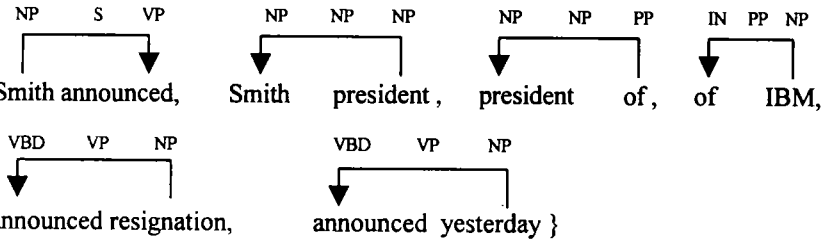
(b) (S (NP (NP John/NNP Smith/NNP),(NP (NP the/DT president/NN)
(PP of/IN IBM/NNP),)
(VP announced/VBD (NP his/PRP$ resignation/NN) yesterday/NN).)

(c)



[John **Smith**] [the **president**] of [**IBM**] announced [his **resignation**] [**yesterday**]

B={[John Smith],[the president],[IBM],[his resignation], [yesterday]}



D={Smith announced, Smith president, president of, of IBM,
announced resignation, announced yesterday }

基于以上模型的分析器的分析过程可以描述如下：对于一个输入句，首先分析出其中的base NP，然后根据训练数据中得到的依存概率计算各个依存关系集合的可能性，可能性最大的依存关系集合就成为句子成分结构的最佳估计。由于依存关系表示为一个三元组，因此依存关系集合和base NP集合就可以映射为句子的短语结构树。

Collins的算法是以大规模树库为基础的，而且以完全句法分析为目标。这种方法以基本名词短语识别为前提，利用具体词之间的依存概率，把短语归结为其中心词。这个研究思路对于部分句法分析也是很有借鉴意义的。

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：(1) 增加句法标记法(incremental/constructive approach)，即在词串中插入短语边界和短语类型等句法标记；(2) 删除句法标记法(reductionist approach)，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(transducers)序列，转换器由正则式(regular expression)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Abney的语块分析系统CASS (Abney 1991, 1996b)， Mokhtar and Chanod (1997)等。下面以Abney的有限状态层叠(finite-state cascades)方法为例对此加以说明。

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T_i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(level recognizer) T_i，它以L_{i-1}级的输出为输入，并产生L_i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。例如，给定下面的规则：

$$T_1: \left\{ \begin{array}{l} NP \rightarrow (D)\ A^*\ N^+ \\ VP \rightarrow V\text{-tns} \mid Aux\ V\text{-ing} \\ NP \rightarrow Pron \end{array} \right\}$$

T_2: { PP →P NP }

T_3: { S →PP* NP PP* VP PP* }

对输入句“the woman in the lab coat thought you were sleeping”的分析过程如下所示：

L_3	----------						S		----------	S	(T_3)
L_2	---------NP		----------		PP		VP	NP	----------	VP	(T_2)
L_1	---------NP	P	----------		NP		VP	NP	----------	VP	(T_1)
L_0	D	N	P	D	N	N	V-tns	Pron	Aux	V-ing	
	the	woman	in	the	lab	coat	thought	you	were	sleeping	
	0	1	2	3	4	5	6	7	8	9	

识别器T_1从第0个词开始，在L_0级上进行匹配，在到达位置2时，得到一个与NP模式相匹

配的状态序列，于是在L_1级上，从位置0到位置2输出一个NP。然后从位置2重新开始，因为没有与之匹配的模式，所以把P直接输出。然后又从位置3开始，在位置5和位置6上分别有一个与NP模式相匹配的模式，这时采用最长匹配，于是在L_1级上从位置3到6输出一个NP，然后又从位置6起继续匹配。

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：

（1）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）；

（2）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。

输入句： others moved away from traditional jazz practice.

经过词性标注后，加上可能的句法标记：

“<others>”	“other”	PRON	@>N	@NH
“<moved>”	“move”	V	@V	
“<away>”	“away”	ADV	@>A	@AH
“<from>”	“from”	PREP	@DUMMY	
“<traditional>”	“traditional”	A	@>N	@N< @NH
“<jazz>”	“jazz”	N	@>N	@NH
“<practice>”	“practice”	N	@>N	@NH

标记注释：@>N（前定语） @N<（后定语） @NH（NP核心） @>A（前状语）
@A<（后状语） @AH（副词短语核心） @V （动词和助动词） @DUMMY（介词）

在上面的例子中，第1列是词语，第2列是词的原形，第3列是词性标记，第4列是句法标记。所有的句法标记都以@开头。如果一个词有两个或两个以上的句法标记，则说明它在句法上是有歧义的。在句法分析过程中根据语法规则进行消歧。如果一个词只有一个标记，则不运用规则，如果一个词虽有歧义的标记，但没有与之匹配的规则则保留歧义。即分析结束后不保证每个词都只有一个句法标记。

规则采用所谓限制语法(constraint grammar)的形式。如：

REMOVE (@>N) (*1C <<< OR (@V) OR (@CS) BARRIER (@NH))

这条规则的含义是：如果上下文满足下面的条件，则从有歧义的词中删去前定语标记@>N：其右边某个词(*1中，*表示一个或多个，1表示右边)是非歧义的(C)，这个词是句子边界(<<<)、动词（@V）或主从连词（@CS），并且该词和当前词之间没有哪个词有@NH标记。

这些规则主要是人工书写的(Voutilainen 1993; Voutilainen and Padro, 1997)，但完全靠人工总

结这些上下文限制规则是十分困难的，可以采用机器学习的方法从语料库中自动获取这些语法规则(Brill 1995; Samuellson *et al.* 1996)。

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。Eric Brill (1995)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。Ramshaw and Marcus(1995)把这种自学习方法用于识别英语中的基本名词短语(base NP)。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(instance-based learning)或基于记忆的方法(memory-based learning)，如 Cardie and Pierce(1998)和 Argamon *et al.*(1998)。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。

3.3.1 基于转换的规则学习方法

如图1所示，基于转换的学习方法以下列三部分资源为基础：(1) 带标注的训练语料库。对于base NP识别任务来说，训练语料要标注出其中所有正确的base NP（在此之前当然要先标注词性)。(2) 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。(3) 一个初始标注程序。

基于转换的错误驱动的学习算法是：

(1) 初始标注。把训练语料中所有的base NP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base NP。把这个结果作为系统的底线(baseline)。

(2) 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动作。

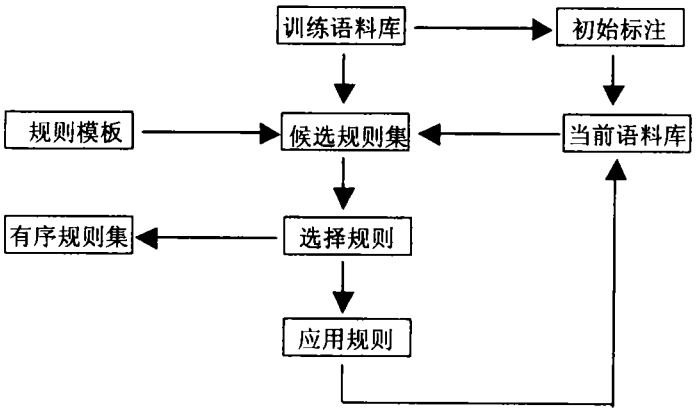


图1 基于转换的错误驱动的学习过程

(3)获取规则。把候选规则集中的每条规则分别运用于初始标注的结果，选出得分最高的规则(得分为正确的修改数减去错误的修改数得到的结果)。把这条规则运用于初始标注的结果作为下一轮循环的基础，并把这条规则作为规则序列中的第一条规则输出。重复以下过程直到得分最高的规

则的得分为0或低于某个阈值为止：获取候选规则集，给其中每条规则打分，选择得分最高的规则输出到规则集中，并把这条规则作用于当前语料库。

通过以上的自动学习过程就可以得到一个有序的规则集。base NP 识别的过程是，首先运用初始标注程序标注出输入句中可能的base NP，然后顺序运用规则集中的规则对初始标注的结果进行转换操作。

3.3.2 基于实例的规则学习方法

前面所介绍的基于转换的学习方法在学习过程之后得到的是识别短语的规则，这样的规则描述在什么条件下一个序列是一个基本名词短语，在什么条件下不是一个基本名词短语。而基于实例的学习方法是通过学习得到一组短语的组成模式，分析的时候利用这样的模式去和文本中的词类序列进行匹配。

Cardie and Pierce（1998）把标注好短语信息的语料库分为两个部分，一部分用于训练，另一部分用于剪枝。首先从训练的语料中得到一组名词短语的组成模式规则，然后把得到的这些规则应用到剪枝的语料中，对这些规则进行打分。比如，如果一个规则识别出一个正确的短语得1分，识别出一个错误的短语得-1分，这样根据每条规则的总的得分情况对规则进行删减，去掉那些得分低的规则。最后得到的一组规则能保证较高的正确率。应用这些规则来识别文本中的名词短语的方法很简单，就是简单的模式匹配方法，在遇到规则冲突时，采用最长匹配原则。

Argamon *et al.*(1998)并不是通过学习过程显性地得到一组短语的组成模式(词类序列及其上下文环境)，这些模式隐含在标注好短语的实例中。在训练阶段，把标注好词性和短语边界的语料用一种可以快捷检索的数据结构存储起来。在识别阶段，拿文本中的词类序列和训练语料中的实例进行匹配：把句子中的每个子串作为候选，对于每个候选，通过查找实例库计算它的概率分数，对分数高于某个阈值的候选予以保留。这一技术的关键在于候选子串和实例的匹配，因为子串可能是若干词的序列，而且还要考虑上下文。如果拿整个子串去和实例匹配的话就有严重的数据稀疏问题，这是基于实例的方法中普遍存在的一个问题。他们提出的一种覆盖(cover)技术较好地解决了这一问题。覆盖技术的基本思想是对待分析的串（包括上下文）进行分解，把它分解成若干更小的子串，利用这些子串去匹配。最后找到一个覆盖原串的子串集合，这些子串的总的概率分数最高。这种基于实例的方法把语法规则隐含在标注好的实例之中，跟前两种学习方法相比，它并没有一套显性的用于识别的语法规则，所以这种方法似乎更像基于统计的方法。

4. 汉语的有关研究

近年来，我国学者也开始借鉴国外的方法进行汉语浅层句法分析的探索。李文捷等(1995)用短语边界与词性标记对共现概率的方法研究汉语中最长名词短语的识别。首先在训练集中统计NP起始和NP终止两个概率矩阵，然后根据这些概率信息在输入句的词性标记对之间插入NP起始标记和NP终止标记，并对标记进行匹配处理。张国煊等(1995)用简单的互信息方法划分短语边界。郭志立等（1996）用互信息方法确定汉语“的”字短语的边界。这些研究都是基于统计的方法的。

孙宏林（1997）用规则方法识别汉语VO（动宾）结构；刘长征（1998）采用规则方法识别由名词序列构成的NP。这两项研究中的语法规则是根据语料库中的统计数据通过人工归纳得到的。赵

军(1998)系统地研究了汉语基本名词短语的识别和分析。在识别方面，从预先定义的句法模板(组成base NP 的词类序列）出发，探讨了两种base NP的识别方法：一种是统计N元模型，该模型利用了base NP 组成成分的词性信息、音节信息及上下文信息，研究表明这种模型比单纯的基于词类序列的模型要好。另一种是规则方法，其规则通过基于转换的学习方法从训练语料中自动获取(赵军、黄昌宁1999)。周强（1996）的研究目标虽然是句子的完整句法分析，但其第一步是短语边界的初界定，利用从树库中得到的统计信息确定一个词前后的边界类型。穗志方(1998)的目标也是完全分析，基本思想是，首先确定谓语中心词，然后围绕谓语中心词进行由底向上的组块分析，以确定谓语中心词的支配成分。其中的组块分析如果独立出来，就是一个部分句法分析器。在这个组块分析过程中主要利用了词语之间的依存关系。

5. 结束语

本文简要介绍了近年来在浅层句法分析领域出现的一些有代表性的方法。由于我们掌握的文献有限，可能会有一些遗漏，但基本上可以看到该领域近年来的基本动向。九十年代以来，自然语言处理领域在技术的发展中出现了两种十分突出的现象：一是统计方法得到了广泛的应用，二是有限状态方法进一步扩大其应用范围。这两种现象都在浅层句法分析领域得到了充分的反映。另外值得注意的是，研究如何采用机器学习方法进行语言知识的自动获取已经成为一个十分明显的趋势。人们常常有这样一种认识，即基于统计的方法是经验主义的，基于规则的方法是理性主义的，而且这两种方法是根本对立的。但如果语法规则都是从语料库中自动或半自动获取的，那么这两种方法还是那样泾渭分明吗？由此可以看到，目前的自然语言处理研究不管是采用统计方法还是采用规则方法，都离不开语料库这一基础资源。宾州大学的英语树库(Marcus *et al.* 1993)极大地推动了英语在这方面的研究，以上介绍的许多研究工作都是以这一语料库为基础的。所以，要想真正提高汉语信息处理的水平，首先必须开发大规模、高质量的标注多种信息的能够共享的汉语语料库。

参考文献

Abney, Steven. 1991. Parsing by chunks. In Robert Berwick, Steven Abney and CaroTenny, eds., *Principle-Based Parsing*. Dordercht: Kluwer Academic Publishers.

----. 1996a. Part-of-speech tagging and partial parsing. In Ken Church, Steve Young , and Gerrit Bloothooft, eds., *Corpus-Based Methods in Language and Speech*. Dordercht: Kluwer Academic Publishers. Pp. 119-136.

----. 1996b. Partial parsing via finite-state cascades. In *Proceedings of the ESSLLI '96 Robust Parsing Workshop*.

Argamon, S., I. Dagon and Y. Krymolowsky. 1998. A memory-based approach to learning shallow natural language patterns. In *Proceedings of COLING-ACL '98*. Pp. 67-73.

Brill, Eric. 1995. Unsupervised learning of Disambiguation Rules for part of speech tagging. In *Proceedings of the 3rd Workshop on Very Large Corpora*. Pp. 1-13.

Cardie, Claire and David Pierce. 1998. Error-driven pruning of treebank grammars for base noun phrase identification. In *Proceedings of COLING-ACL '98*. Pp. 218-224.

Chen, Kuang-hua and Chen, Hsin-Hsi. 1994. Extracting noun phrases from large-scale texts: a hybrid approach and its automatic evaluation. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*. Pp. 234-241.

Chen, Hsin-Hsi and Lee, Yue-Shi. 1995. Development of a partially bracketed corpus with part-of- speech information only. In *Proceedings of the 3rd Workshop on Very Large Corpora*. Pp. 162-172.

Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*. Pp. 136-143.

Collins, M. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*. Pp.184-191.

Fano, R. M. 1961. *Transmission of Information, A Statistical Theory of Communication*. MIT Press.

Gale, W. A. and K. W. Church. 1991. Identifying word correspondences in parallel texts. In *Proceedings of DARPA Speech and Natural Language Workshop*. Pp.152-157.

Magerman, D. and M. Marcus. 1990. Parsing a natural language using mutual information statistics. In *Proceedings of AAAI '90*, Pp. 984-989.

Magerman, D. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the ACL*.

Marcus, M., B. Santorini, and M. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19, 2: 313-330.

Mokhtar, S. and J. Chanod. 1997. Incremental finite-state parsing. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. Pp.72-79.

Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In Morgan Kaufmann, Waibel and Lee, eds., 1990, *Readings in Speech Recognition*. Pp. 267-296.

Ramshaw L. and M. Marcus. 1995. Text chunking using transformation-based learning. In *Proceedings of the 3rd Workshop on Very Large Corpora*.

Samuelsson, C., P. Tapanainen and A. Voutilainen. 1996. Inducing constraint grammars. In *Grammatical Inference: Learning Syntax from Sentences*. Springer-Verlag. Also available at cmp-lg. 9607002.

Skut,Wojciech and Thorsten Brants. 1998. A maximum-entropy partial parser for unrestricted text. In *Proceedings of the 6th Workshop on Very Large Corpora*. Montreal, Quebec, 1998. Also available at cmp-lg. 9807006.

Voutilainen, A. 1993. Nptool, a detector of English noun phrases. In *Proceedings of the First Workshop on Very Large Corpora*.

Voutilainen, A. and L. Padro. 1997. Developing a hybrid NP parser. In *Proceedings of the 5th Conference on Applied Natural Language*, Pp. 80-87.

郭志立、苑春法、黄昌宁，1996，用统计方法研究“的”字短语的结构与边界。《计算机时代的汉语和汉字研究》（罗振声、袁毓林主编）北京：清华大学出版社。

刘长征，1998，基于词性标注语料库中普通名词序列的捆绑研究。《1998 中文信息处理国际会议论文集》（黄昌宁主编）清华大学出版社。

李文捷、周明等，1995，基于语料库的中文最长名词短语的自动抽取。《计算语言学进展与应用》（陈力为、袁琦主编）北京：清华大学出版社。

穗志方，1998，语句相似度研究中的骨架依存分析法及其应用。北京大学计算机系博士论文。

孙宏林，1997，从标注语料库中归纳语法规则：“V+N”序列实验分析。《语言工程》（陈力为、袁琦主编）北京：清华大学出版社。

张国煊、郁梅、王小华，1995，基于语料库的汉语边界划分的研究。《计算语言学进展与应用》（陈力为、袁琦主编）北京：清华大学出版社。

赵军，1998，汉语基本名词短语识别及结构分析。清华大学计算机系博士论文。

赵军、黄昌宁，1999，基于转换的汉语基本名词短语识别模型。《中文信息学报》第 13 卷第 2 期。

周强，1996，一个短语自动定界模型。《软件学报》第 7 卷增刊。

作者通讯地址：　孙宏林　北京语言文化大学语言信息处理研究所　邮编 100083
　　俞士汶　北京大学计算语言学研究所　邮编 100871

《当代语言学》第2卷2000年第2期124页，北京

Abstracts of Articles

Zhan Weidong, A survey of Chinese information processing since 1980

This paper presents a critical review of the research in the field of Chinese information processing since 1980. It falls into three categories: (1) studies introducing theories and methods by overseas researchers; (2) studies exploring new approaches to Chinese information processing; and (3) the construction of Chinese linguistic knowledge databases. It is shown that, among the studies of contemporary Chinese grammar facilitating Chinese information processing, the research into formalized phrasal structure rules appears most promising.

Sun Honglin, Shallow parsing: an overview

This paper presents an introductory overview of the studies on shallow parsing for both English and Chinese. The major methods within two paradigms, statistics-based and rule-based, are discussed in detail. Statistics-based methods include HMM-based method, mutual information, chi-square statistics, and dependency probability of heads. Rule-based methods include incremental approach, reductionalist approach, and the methods for learning rules.

Feng Zhiwei, Some approaches to automatic parsing based on phrase structure grammar

PSG (Phrase Structure Grammar) has been extensively applied in computational linguistics. In this paper, some automatic parsing approaches based on PSG were introductorily surveyed. They are: top-down parsing method, bottom-up parsing method, Tomita algorithm, left-corner parsing method, and CYK algorithm.

Luo Qijing, A quick review of the Chinese translation of Masini's *The Formation of Modern Chinese Lexicon and Its Evolution Towards a National Language: The Period from 1840-1898*

This quick review attempts to show that, contrary to a HK reviewer's claim that the Chinese translation of Masini's work is a phenomenal achievement, the translation is in fact phenomenally poor. A preliminary reading of it quickly spots ten translation errors, including things such as misunderstanding of word meanings and sentence structure in both the source language (English) and the target language (Chinese). For example, the translator renders "since the Canton days" as 自广州那时起。For another example, he is sometimes unaware of some common background knowledge: 八股文 (a style, i.e. the so-called "eight-legged essay") is translated as 文言 (Chinese classical writing). All these mistakes can be attributed to the translator's lack of linguistic competence in both English and Chinese.