

# 基于 DOP 的汉语句法结构树分析研究\*

郭海旭 邢欣 李福民 李茹

山西大学计算机与信息技术学院, 山西 太原 030006

**摘要:** 本文基于面向数据分析技术 (DOP) 对汉语句子进行了句法结构树分析, 其中, 对于输入句子, 首先需要经过词汇层与词性层初选得到句法片段; 然后, 基于已构建的数据库, 对词汇词性序列的子序列搜索片段, 进行片段组合。最后, 对输入句子与初选结果进行相似性评估, 完成输入句子的组合分析过程。在 CIPS-ParsEval-2009 提供的 task5 语料库上对陈述句事件描述单元进行了训练和测试, 取得的 F-1 测度分别为: Without-head match F1: 72.78%; Complete-head match F1:65%。在 HIT 提供的关于 LOC 类 1252 条进行实验。Close 测试的正确率达到 94%。

**关键字:** 面向数据分析技术、句法片段、句法截断、句型、片段相似、句法树相似

## Research on Chinese Syntax Structure-Tree

### Based on DOP

Guo Hai-xu, Xing Xin, Li Fu-min, Li Ru

School of Computer & Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China

E-mail: haixuguo@126.com

**Abstract:** The paper studies Chinese syntax structure-tree analysis based on data oriented parsing (DOP). At first, every input sentence must be primary selected based on vocabulary level and lexical-category level to get syntax-fragment; then it executes search for database based on subsequence of vocabulary level and lexical category level, and combinations syntax-fragment. Finally, the similarity estimate between the input sentence and the primary selection result is proceeded by using the similarity-based probability estimate technique. So the combination parsing process of the input sentence can be completed successfully. Training set and test set of declarative sentence are based on Task5 database source which is derived from CIPS-ParsEval-2009, the experiment results show that Without-head match F1: 72.78% and Complete-head match F1:74.65%. In addition, training set and test set of question sentence are based on 1252 LOC-database source which is derived from HIT, close-test results show that accuracy rate is 94%.

**Keywords:** Data Oriented Parsing (DOP)、Syntax-Fragment、Syntax-Cut、Sentence Pattern、Fragment Similar、Syntax-Tree Similar

目前, 英语的句法结构分析理论以及技术研究已日趋完善。其中, 完全句法分析中以 (HPSG)<sup>[1]</sup> 分析理论较广泛, HPSG 理论把语言的句法、语义和语用信息利用复杂特征表现在词汇中, 并通过语法规则来约束这些关系, 从而达到描写一种语言的目的。穗志方等<sup>[2]</sup> 利用信息论中熵与条

---

\***基金项目:** 国家 863 高技术研究发展计划资助项目 (2006AA01Z142) 国家社会科学基金青年项目 (07CYY022)

件熵的度量来显示一个特征类型是否抓住了预测句法结构的主要信息。并提出了一种有效的特征选取方法。段湘煜等<sup>[3]</sup>针对决策式依存语法提出了基于动作建模的概率分析算法。实验表明，此方法明显改善了依存树分析性能。在浅层句法分析方面，Abney 提出了块理论并开发了多层次的有限状态成分组块自动识别工具<sup>[4]</sup>。

汉语句法结构分析中，前人曾多次基于规则的方法如：移进—归约（Shift-Reduce Parsing）和图分析（Chart Parsing）技术等来研究汉语句法结构，然而，由于汉语没有严格意义上的形态标志和形态变化，句法规则很难穷尽，而且规则方法对汉语真实语料的处理能力不够。近几年来，随着语料库语言学的不断发展和标注语料库规模的不断扩大，许多研究人员开始尝试利用语料库中的标注信息进行句法分析。如：文献中采用的模拟退火

（Simulated）模型<sup>[5]</sup>和 David M. Magerman 的概率型判定树（Statistical Decision Model）模型<sup>[6]</sup>以及 R. Bod 提出的面向数据分析（Data Oriented Parsing, DOP）<sup>[7]</sup>技术。

Shca 在 90 年代首次提出了面向数据分析（Data Oriented Parsing, DOP）技术<sup>[7]</sup>，DOP 模型首先需要构建一个经过标注的语料库，然后从这个语料库中抽取所有任意大小规模和复杂结构的片段。其次，通过对语料库中片段组合操作来实现新输入的分析，最后考虑输入的所有派生结果的概率总和的大小来选择最有可能性的分析结果。之后国外的 Rod 进行了大量的研究，国内的朱靖波、姚天顺在文章<sup>[8]</sup>首次将 DOP 用于汉语句法结构分析并提出基于分解算法实现句法片段的获得、随后张亮基于 DOP 和中文问句结构研究了中文问句句法结构<sup>[9]</sup>并提出了一种基于句法截断有效快速获得片段的方法，实验结果表明：基于 DOP 进行汉语句子结构分析最大的优点是可行性、可扩展性以及有后续发展空间，它建立在包含大量标注语料库基础上的经验主义方法。随着语料库规模的越来越大，其对各种句子句型的覆盖面必然越来越大，从而其分析的精确率和召回率也必然有所提高，其性能不会达到某个百分点就停止提高。

## 1 DOP 句法分析技术实现

### 1.1 构建数据库

根据 DOP 理论，句法树库与句法片段库是两个重要的数据库，其中，通过对语料库中的句子进行人工加工，构建带有句法标记的句法分析树库。而为了实现片段库的自动构建，我们设计了基于截断的片段自动抽取算法和片段组合自动抽取算法。

在构建语料库之前，先给出句法片段和句法截断的定义。

定义 1.1 句法片段的定义<sup>[9]</sup>

句法片段：从一棵句法树 T 中抽出来的一颗子树 t 称之为 T 用来分析的片段，且 t 具有如下特征：(1) t 的结点个数至少两个；(2) t 中任何一个非叶子结点必须保持与母树中相应结点具有相同个数和关系的子结点；t 是连通的。

以“财政一词源于拉丁文 F i n i s。”为例，如图 1.1：(1) (2) 是正确片段。(3) 是错误片段。

定义 1.2 句法截断的定义<sup>[9]</sup>

句法截断：一个截断是树上全部结点 D 的一个真子集 C，使得：(1) C 中没有一个结点处在由 C 中其他一个结点开始的任何一条后继结点路径上；(2) D 中没有一个结点可以加入 D 而不违背规则。以“财政一词源于拉丁文 F i n i s。”为例，如图 1.1：(2) 是正确截断。(1) 是错误截断。

以“[dj [np 财政/n [np 一/m 词/n ] ] [vp 源于/v [np 拉丁文/n F i n i s

“/n ] ] ]”为例，将部分截断和片段以及片段组合对照如下表 1.1

表 1.1 截断、片段以及片段组合对照表（部分）

截断	片段	片段组合
np vp	(1) [dj [np ] [vp ] ]	(1)+(2)+(3)
	(2) [np 财政/n [np 一/m 词/n ] ]	
	(3) [vp 源于/v [np 拉丁文/n F i n i s/n ] ]	
n 一词 vp	(1) [dj [np /n [np 一/m 词/n ] ] [vp ] ]	(1)+(2)+(3)
	(2) 财政/n	
	(3) [vp 源于/v [np 拉丁文/n F i n i s/n ] ]	

通过分析，发现所有的句法片段都可以由句法截断生成，即片段或者是整个句法树的某个截断的上部或下部(含截断)，或是句法树的子树(节点数须大于 2)的截断的上部或下部(含截断)。而且截断有助于后面问句句型统计算法实现。通过截断看片段，使得片段及其片段组合的概念更加明晰。

## 1.2 DOP 句法分析步骤

### 1.2.1 处理流程

处理流程见图 1.2

### 1.2.2 具体步骤

#### ① 对输入句子进行分词和词性标注

采用山西大学自己研发的词法分析器 F2000 进行分词、词性标注。

#### ② 将汉语句子分隔成多个事件描述单元

按照 CIPS-ParsEval-2009<sup>[12]</sup>中事件描述单元的定义，设计算法实现事件描述单元的自动识别。

#### ③ 基于词汇序列与词性序列匹配

如果语料库中某句的词汇序列和输入句子的词汇序列完全匹配，并且词性序列和输入句子的词性序列完全匹配，则输出该句对应的句法树结构，结束以下分析。否则转④

#### ④ 判断句子是否有疑问词，若有转⑤，否则转⑦

#### ⑤ 基于疑问词位置的初选

问句疑问词位置有句首、句中、句尾三种分布。然后从树库中寻找与输入问句的疑问词位置一致的问句。

#### ⑥ 基于问句句型与词性序列的匹配

若语料库中某句的句型全部是由词汇和词性组成的序列，并且两句的词性序列也完全相同，则可将语料库中此句的句法树作为输入问句的句法树，只是把语料库中句子的词汇序列换为输入问句的词汇序列，即可得

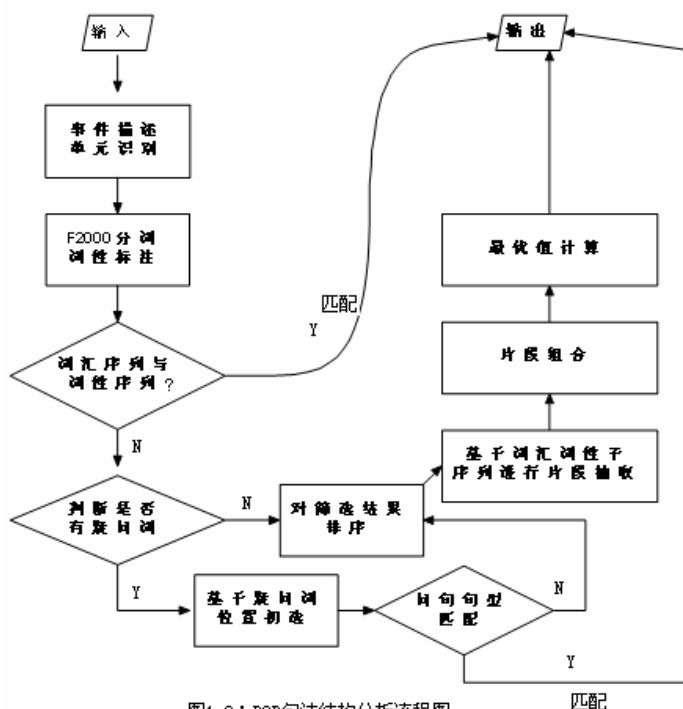


图1.2：DOP句法结构分析流程图

到输入问句的句法树，结束以下分析；否则搜索与输入问句的词性词汇序列的子序列最大匹配的句型，然后进入下一步。（句型最大匹配详见 1.3）

⑦ 对筛选结果排序

如果输入的句子有疑问词，则利用⑤筛选出的问句的词性序列与输入句子的词性序列比较；否则用树库中的所有句子的词性序列与输入句子的词性序列比较。计算两者具有相同词性子串的总长度。记相同子串序列为  $S=S_1\hat{S}_2\hat{\dots}S_k$ ，输入问句词性序列为  $F=f_1\hat{f}_2\hat{\dots}f_k$ ，比较句子词性序列为  $W=W_1\hat{W}_2\hat{\dots}W_k$ ， $p(S, F, W)=\min(\text{long}(F), \text{long}(W))-\text{long}(S)$ ；其中函数  $\text{long}(x)$  为词串  $x$  的长度。初选按  $p(S, F, W)$  值由小到大的次序，将语料库中的句子排列。再确定一阈值，当  $p(S, F, W)$  大于阈值时，便不再考虑此句子。遇到两个  $p(S, F, W)$  相同的情况按照词汇相同的个数排序。（实验阈值设置为 2）

⑧ 基于词汇和词性序列的子序列搜索

检索片段库，搜索以输入语句中的任一词性串（或者词汇串）作为叶结点的片段。

⑨ 片段组合处理

利用⑦得到的句子，抽取其片段并与⑧得到的片段进行组合，搜索能够组合成叶子节点为输入句子中各组成词汇的句法树。

⑩ 最优值计算

由于输入句子可能有多种组合形式，所以要通过最优值计算来综合考虑各种因素，筛选出最有可能的答案。基于输入句子的各种组合形式，与已排序的初选结果中的句子进行相似度计算，这里我们只与第⑥步得到的句子的片段组合进行相似度计算。

1.3 问句句型统计和分析

著名语言学家吕书湘先生曾经指出：“怎样用有限的格式去说明繁简有方，变化无穷的句子，这应该是语法分析的最终目的，也应该是对于学习的人更为有用的工作。”<sup>[13]</sup>首先，对问句的训练语料进行问句句型统计。80 年代以来，许多语言学家对句型问题进行了大量的研究，他们从语法、语义、语用三个不同的层面探索，提出了不同的句型分类标准和系统。在国内有关句型研究时间还不长，其中，北京语言大学句型研究小组从对外汉语教学需要出发提出了一个句型系统；清华大学人文学院计算语言学研究室以汉语句型的自动分析为基础，实现了一个汉语举行的频度统计算法，在不包括疑问句的情况下统计归纳出了 191 种句型<sup>[14]</sup>。张亮<sup>[7]</sup>对问句句型进行了形式定义，他认为一个问句句型就是前面讲过的一个截断，如果一个截断是一个句型，当且仅当它在训练库中涵盖的问句数量大于一个阈值。阈值越大，句型越少，代表性越强。

本实验基于张亮的问句句型定义，考虑到每个句型在训练库中的分布比例，同时考虑到每个句型中的每个结点在句法结构中的层次。按照如下的权重公式<sup>[7]</sup>进行统计

$$Q(c) = \log(f_c) \times \sum_{i=1}^m (\text{deep}(x_i) + \alpha \times \text{ques}(x_i) + \beta \times \text{word}(x_i))$$

其中  $f_c$  是句型  $c$  在训练库中出现的频率。  $\text{deep}(x_i)$  是结点  $x_i$  在句法结构中的层次。  $M$  是句型  $C$  的句法结构中结点的数目，  $\alpha$  和  $\beta$  分别是计算词汇系数和疑问词系数，在

$$\text{word}(x_i) = \begin{cases} 1 & \text{当结点 } x_i \text{ 为问句中的词汇时} \\ 0 & \text{否则} \end{cases}$$

$$\text{ques}(x_i) = \begin{cases} 1 & \text{当结点 } x_i \text{ 为问句中的疑问词时} \\ 0 & \text{否则} \end{cases} \dots\dots\dots \alpha=25, \beta=20$$

算法中取 25 和 20，在问句中的疑问词是问句的重要信息特征，对句型结构有很重要的贡献，因

此计算中需突出疑问词的作用。本试验共 66 个问句，其中问句句型 14 个，疑问词 33 个。

对输入经过正确分词和词性标注的问句如何进行句型匹配呢？

我们利用输入问句的词汇和词性序列去句型库中寻找最大的匹配子串。

例如：我们输入的是“山西/nsh 五台山/ns 位于/v 哪儿/r ? /w”，我们搜索这样的句型“np+v+哪儿”，然后再到片段库中搜索[np /nsh /ns ]和[v 位于]句法片段。实现句法结构的分析。如果这样的片段都能搜索到，结束分析；否则进入下一步。

## 2 相似度概率评估技术

本实验基于 Kullback-Leibler 距离的距离函数实现相似度评估技术，通过距离函数计算不通概率分布之间的距离，距离越小，相似程度越大。

### 2.1 概率分布评估

对象  $a$  的条件概率分布  $P^a$  使用最大似然估计方法 (Maximum Likelihood Estimate, MLE) 来计算。

MLE 方法使用训练数据来评估下面的条件概率  $P^a = P_{MLE}(c/a)$  :

$$P_{MLE}(c/a) = \frac{R(a,c)}{R(a)} \quad \text{其中, } R(a,c) \text{ 为对象 } a \text{ 在局部上下文 } c \text{ 条件下在训练语料中出现的次数, } R(a) \text{ 表示对象 } a \text{ 在训练语料中出现的次数。}$$

由于语料规模的限制，如果某一数据对未出现在样本语料中，则得出其条件概率为 0，许多数据对即使出现在训练样本语料中，计算得到的结果也是低概率事件，这就是稀疏数据的问题。眼中的稀疏数据会影响评估的准确率，为此，研究者们提出了 MLE 的改进方法。其基本思想是：首先将 MLE 作为最初的评估，然后将数据对的条件概率之和小于 1，留出一部分概率赋给未出现在样本语料中的可能数据对，即平滑 (Smoothing) 技术。Jelinek 和 Mercer 提出了内插法平滑技术。其简化的内插法平滑公式为：

$$P_{MLE}(c/a) = \lambda \frac{R(a,c)}{R(a)} + (1-\lambda) \frac{1}{|CT|} \quad \text{其中, } |CT| \text{ 表示局部上下文的数目。在本次试验中,}$$

对象  $a$  相当于句法树中的某个节点。局部上下文采用该节点的父辈节点 (句法标记，如 NP, VP 等)。设置插值参数为 0.99。

### 2.2 基于 Kullback-Leibler 距离的距离函数

定义  $O_1$  和  $O_2$  为两个不同的对象， $CT$  为局部上下文， $c \in CT$  表示某一具体的局部上下文。 $P_{O_1}$  和  $P_{O_2}$  分别表示对象  $O_1$  和  $O_2$  的概率分布。Kullback-Leibler 距离公式定义如下：

$D(P_{O_1} || P_{O_2}) = \sum_{c \in CT} P(c/O_1) \log \frac{P(c/O_1)}{P(c/O_2)}$  可以利用  $D(P_{O_1} || P_{O_2}) + D(P_{O_2} || P_{O_1})$  来计算两个不同的概率分布  $P_{O_1}$  和  $P_{O_2}$  的距离。

即  $\xi(P_{O_1}, P_{O_2}) = D(P_{O_1} || P_{O_2}) + D(P_{O_2} || P_{O_1})$  该距离函数用于句法树片段间的相似性的判定过程。

### 2.3 句法片段相似度和句法树相似度

进行相似度比较的时候我们要考虑这样几个因素：

A. 根节点的相似性； B: 相同结点的个数； C: 相同结点在片段中的层次； D: 是否含有疑问

词及疑问词是否相等。

公式一：片段相似度公式<sup>[7]</sup>

$$Sim(x, y) = \frac{2 \times A}{B} \times (1 + \theta) \times e^{-\xi(P_{root}(x), P_{root}(y))}$$

其中：

$$A = \sum_{k=1}^n (Layer(same(x, y)_k)), B = \sum_{i=1}^m (layer(x_i)) + \sum_{j=1}^h (Layer(y_j))$$

$$\theta = \begin{cases} 0 & \text{当 } x \text{ 与 } y \text{ 都不含疑问词} \\ \eta & \text{当 } x \text{ 与 } y \text{ 含相同疑问词} \\ -\eta & \text{其它} \end{cases} \quad 0 \leq \eta \leq 1$$

$Same(x, y)$  : 片段  $x$  和  $y$  中相同组成结构的对应结点。

$Layer(z)$  表示结点  $z$  的层次； $n$  为相同节点的个数， $m$  是片段  $x$  中的个数， $h$  是片段  $y$  的个数； $root(x)$ ：片段  $x$  的根节点。当  $root(x) = root(y)$  时，

$\xi(P_{root}(x), P_{root}(y)) = 0$ 。当两个片段完全相同时，则  $Sim(x, y) = 1$ 。实验中  $\eta = 0.2$ 。

公式二：句子相似度公式<sup>[7]</sup>：

其中： $T_X$  是句子  $X$  的句法树；

$FR(T_X) = \{x_1, x_2 \dots x_i, \dots x_n\}$  是  $T_X$  的片段组合形式； $x$  是片段； $n$ 、 $m$  分别是句子  $X$  和  $Y$  的片段组合形式的个数。

$$\begin{aligned} Sim(X, Y) &= Sim(T_X, T_Y) \\ &= \arg \max (Sim(FR(T_X)_i, FR(T_Y)_j)) \\ &= \arg \max (Sim(x_i, y_j)) \\ &= \arg \max (\prod_{k=1}^n Sim(x_i^k, y_j^k)) \end{aligned}$$

根据输入句子与初选结果之间的相似度计算，当相似度第一次大于所设定的阈值时，即可将此时输入句子所对应的片段组合形式作为正确的分析结果。如果完成输入句子与初选结果之间的相似度计算后，依然不存在大于阈值的相似度。这种情形的解决途径是，将相似度值最高时，输入句子所具有的片段组合形式作为分析结果输出。

### 3 实验结果与实验分析

#### 3.1 实验一：基于 DOP 的陈述句事件描述单元的句法结构分析

实验语料采用的是清华大学信息技术研究院语音和语言技术中心开发的清华汉语句法树库 (Tsinghua Chinese Treebank ver1.0, TCT Ver10)资源。其中训练语料一共 171 篇文档：BAIKE(23 篇)、NEWS(125 篇)、HYL(23 篇)；句子数 14248 条；事件描述单元有 67174 条；测试语料一共 43 篇文档：句子数 3751 条；事件描述单元有 16210 条。

表 3.1：训练语料与测试语料中事件描述单元词数分布

词数		1	2-6	7-10	11-12	13-15	16-20	20 以上	词数最多
训练语料	Count	34551	10807	8724	3165	3456	3257	3214	137 个词
	Percentage	51.4%	16.1%	13%	4.7%	5.1%	4.9%	4.8%	
		51.4%	33.8%			14.8			
测试语料	Count	8271	2642	2215	757	834	785	706	108 个词
	Percentage	51%	16.3%	13.7%	4.7%	5.1%	4.8%	4.4%	
		51%	34.7%			14.3%			

从表中可得：训练语料与测试语料中约 85% 的事件描述单元的词数  $\leq 12$ ，在实验中我们发现，词数对 DOP 句法分析的影响很大，尤其表现在分析时间慢。为此，我们本次试验仅仅针对词数

<=12 个词的事件描述单元进行句法结构分析。

实验的评测按照 CIPS-ParsEval-2009<sup>[12]</sup>中的评价指标：F-1 测度 (F-1 measure),结果如表 3.2

表 3.2: 1-12 个词的 13885 条事件描述单元分析结果

Word-Count (1-12)	Without-head match F1		Partial-head match F1		Complete-head match F1	
	correct	error	correct	error	correct	error
Percentage	10105	3780	9545	4340	9024	4861
	72.78%		68.74%		65%	

### 3.2 实验二：基于 DOP 的疑问句事件描述单元的句法结构分析

实验语料采用的是哈工大信息检索研究室问答系统问题集 trainquestion\_for XML\_20060717.xml 中的 882 个 LOC (地点类) 问句作为训练集, testquestion\_for XML\_20060717.xml 中的 370 个 LOC (地点类) 问句作为测试集。

表 3.3: 疑问句实验数据分析结果

类 型	完全匹配句 子数	句型匹配句 子数	片段组合后输 出	分析错误	平均分析时间 (秒/条)
Count	0	175	173	22	3
Percentage	0.94			0.36	

### 3.3 实验分析

实验表明: 疑问句句法分析的效果明显高于陈述句句法分析效果, 主要是因为疑问句形式上具有一定的相似结构, 有 47% 的结果由问句句型直接得到了句法树, 然而陈述句根据句型几乎得不到完整的句法树; 另外, 在对陈述句句法结构分析实验时, 事件描述单元的词数直接影响分析时间和准确率。在评测中, 中心语的确定上没能将语义和语境加进去, 为此, 有些句法树正确了, 但中心语位置确定错了。

实验中, 未对较长的事件描述单元做处理, 但是我们发现, 可以考虑句群来处理较长事件描述单元。是我们下一步的工作

## 4 下一步工作

本文主要介绍了在基于 DOP 中文句子句法分析方面所做的一些研究工作, 针对疑问句和陈述句分别进行了实验。下一步的工作是:

- ① 基于 DOP 进行山西旅游问句句法结构分析以及山西旅游问句语义结构分析;
- ② 针对词数较多的句子按照句群进行句法结构分析。

致谢: 我们来自刘开瑛教授牵头的基于 CFN 句法语义研究团队, 参赛者有李茹教授和五个学生 (郭海旭、李福民、邢欣、郭鹏奎、温锋瑞), 在评测中, 受到刘开瑛教授的悉心指导; 另外, 在实验中, 得到了 HIT 提供的问句句料, 在此, 一并向他们表示诚挚的感谢。

### 参考文献 :

- [1] Pollard,C.and I.A.Sag. Head\_Driven Phrase Structure Grammar.Chicago:The University of Chicago Press. 1994.
- [2] 穗志方,赵军,俞士汶.统计句法分析建模中基于信息论的特征类型分析,《计算机学报》, 2001 年, 第 24 卷第 2 期
- [3]Xiangyu Duan,Jun Zhao,Probabilistic Parsing Action Models for Mul-ti-lingual Dependency Parsing.In Proceedings of CoNLL Shared Task Session of the Conference on Empirical Methods in Natural Language

Processing and Computational language Learning(EMNLP-CoNLL),June 28-30,2007,Prague,Czech.

[4]Abney,S.Partial parsing via finite-state cascades. Natural Language Engineering ,1996,2(4):337-334.

[5] T.Winograd.Language as a cognitive process. Vol1, Syntax, 116-129.1983.10.

[6]David M. Magerman. Natural Language Parsing as Statistical Pattern Recognition, Doctoral dissertation, Stanford University, Stanford, USA.1994.1.

[7] Rens Bod.A Computational Model of Language Performance: Data Oriented Parsing, In Proc. of COLING-92, Nantes.1990.3.

[8] 朱靖波, 姚天顺.面向数据的句法分析技术.中文信息学报.1998.12.

[9] 张亮,王树梅,黄河燕.面向中文问答系统问句句法分析.山东大学学报.2006.4.

[10]周强.汉语基本快描述体系 中文信息学报 2007 第 3 期

[11]周强, 任海波, 詹卫东(2001)。“构建大规模汉语语块库”, 黄昌宁, 张普生编《自然语言理解与机器翻译》2001

[12] <http://www.ncmmmc.org/CIPS-ParsEval-2009/index.asp>

[13]吕叔湘, 汉语语法论文.北京:商务印书馆, 1984.481-570,533.

[14]罗振声,郑碧霞.汉语句型自动分析和分布统计算法与策略的研究.中文信息学报.VOL.8.No.2.1993.6..