# 汉语结构优先关系的自动获取*

周 强 黄昌宁

(清华大学计算机科学与技术系 北京 100084)
(清华大学智能技术与系统国家重点实验室 北京 100084)
E-mail: zhouq@s1000e.cs.tsinghua.edu.cn

**摘要** 提出了一种用于定量描述歧义结构分布特点的知识表示机制:结构优先关系SPR(structure preference relation),介绍了针对不同语料文本的两种SPR获取方法:树库语料的自动发现和原始文本的自动获取,并且通过不同的实验证明了这些方法的可行性和实用性.另外,还介绍了SPR的若干应用前景,并提出进一步的研究方向.

**关键词** 结构优先关系,歧义结构,知识获取,语料库.

**中图法分类号** TP18

歧义结构的识别和处理是自然语言理解的重要研究课题.针对英语中最常见的歧义结构:介词短语连结(Prepositional Phrase Attachment)问题,许多研究人员进行了不同排歧方法的探索.如:Hindle和Rooth[1]利用从自动分析的语料库中获取的介词与中心动词以及中心名词的不同关联强度来排除歧义.R. Basili等人进一步利用了语义标记信息,提高了排歧能力[2].而Collins和Brooks[3]则采用了Backing-Off方法对此问题进行了处理.由于抓住了英语句法分析中的重点和难点,因而取得了事半功倍的效果.

关于汉语的句法歧义结构,朱德熙先生最早进行了研究,提出了一个有名的歧义结构实例:"咬死了猎人的狗"[4].在此之后,许多语言学家又发现了大量有趣的歧义结构实例.黄国营[5]对这些研究成果进行了概括和总结,形成了一个常见的汉语歧义短语表.但这些研究基本上还局限于结构的枚举和实例的罗列,缺乏比较客观的定量分布数据,如:在真实语料中,各个歧义结构出现的频度有多大?哪种歧义结构最为常见?对于某个特定的歧义结构,各种歧义组合出现的可能性又有多大等等.因此,中文信息处理的研究人员往往需要为大量可能的歧义结构寻找有效的排歧方法,而实际效果常常是事倍功半.

本文就是希望对汉语歧义结构的定量分析方法进行一些探索.它将汉语歧义结构的定量信息有效地集成入结构优先关系SPR(structure preference relation)的描述项中,然后通过对真实文本句子的句法分析树的完全遍历,发现所有可能的SPR实例.而汉语自动句法分析器和句法规则概率参数自动训练算法的开发成功和有效应用,又使得从语料原始句子中自动获取SPR的信息成为可能.目前,初步实验结果证明了这种方法的可行性和实用性.

## 1 结构优先关系的基本内容

**定义** 1. 词类标记集是由所有词类标记组成的集合,简记为POST.

**定义** 2. 句法标记集是由所有句法标记组成的集合,简记为SynT.

**定义** 3. 规则右部是句法规则集中表征合理的句法结构组合的标记串,因其一般在规则描述体系中出现在规则右部,故有此定义.简记为RHP,且有 $RHP \in \{POST \cup SynT\}^*$.

* 本文研究得到国家自然科学基金和中国博士后科学基金资助.作者**周强**,1967年生,博士,助理研究员,主要研究领域为计算语言学.**黄昌宁**,1937年生,教授,博士生导师,主要研究领域为计算语言学.

本文通讯联系人:周强,北京 100084,清华大学计算机科学与技术系

本文1997-10-22收到原稿,1998-02-27收到修改稿

**定义 4.** 交段成分是一种可以在 RHP 的首部和尾部同时出现的成分标记，简记为 IC. 若它处在 RHP 的首部，则 RHP 的其余结构统称为交段后境，简记为 SufIC. 若它处在 RHP 的尾部，则 RHP 的其余结构统称为交段前境，简记为 PreIC. 其中 $IC \in \{POST \cup SynT\}$，$PreIC$，$SufIC \in \{POST \cup SynT\}$.

例如，若规则集中存在这样两条句法规则：$P1 \rightarrow AB$ 和 $P2 \rightarrow BC$，根据其 RHP 内容，可以得到一个交段成分 $B$ 和交段前境 $A$ 及交段后境 $C$.

**定义 5.** 交段结构是由交段前境、交段成分和交段后境共同组成的标记串，简记为 $IS$，即 $IS = \{PreIC\ IC\ SufIC,\ ICPos\}$，其中 $ICPos$ 表示交段成分 $IC$ 在交段结构 $IS$ 中的位置.

例如，$IS = \{v\ v\ n, 1\}$* 就是一个常见的交段结构.

**定义 6.** 交段左向组合实例描述了在分析树中，交段结构中的交段成分首先和左边的交段前境组合而形成的结构实例；而交段右向组合实例则描述了其中的交段成分首先与右边的交段后境组合而形成的结构实例.

**定义 7.** 结构优先关系(SPR)是一个三元组 $\langle IS, LF, RF \rangle$，其中 $IS$ 是交段结构，$LF$ 和 $RF$ 分别表示交段左向和右向组合实例在真实语料中出现的频度.

根据 SPR 中的交段结构在分析树中的组合方式的不同，可以把它分成两大类.

(1) 紧合结构 SPR

SPR 中的交段结构最终组合形成分析树中的一个句法成分，如图 1 中的情况(a)和(c)，其中图 1(a)为交段左向组合实例，图 1(c)为交段右向组合实例.

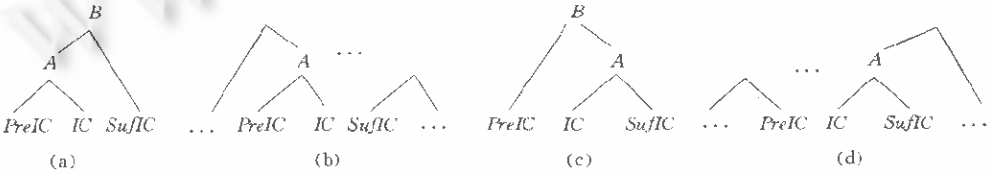


图1 结构优先关系实例的分布特点

(2) 松散结构 SPR

SPR 中的交段结构的交段前境、交段成分和交段后境分别与其相邻左成分或右成分组合成新的句法成分，如图 1 的情况(b)和(d)，其中图 1(b)为交段左向组合实例，图 1(d)为交段右向组合实例.

SPR 的这些结构特点为从分析树中自动发现 SPR 信息提供了方向.

## 2 分析树的 SPR 发现算法

如果存在树库语料，即语料文本的每个句子都标注了正确的句法结构树，则 SPR 的发现就变得很简单了，只需遍历分析树，在每个非叶子节点上发现可能的交段结构实例，再检查 SPR 交段结构的合理性并设置其频度信息.

对一棵分析树 $T$ 的遍历是从它的根节点 $R$ 开始，自顶向下进行的. 其基本处理函数为：

分析树遍历函数 TreeSearch($R$)：

如果树节点 $R$ 是非叶子节点，则

发现紧合结构 SPR；

发现松散结构 SPR；

对每个子节点 $N_i$，循环调用 TreeSearch($N_i$)；

否则，返回.

### 2.1 紧合结构 SPR 的发现

考虑分析树中的子树结构 $PH(RP_1\ RP_2 \ldots RP_n)$，其中 $PH$ 为子树根节点，$RP_i, i \in [1,n]$ 为其子节点成分. 如果 $RP_1$ 也是一个非叶子节点，且其子树结构为 $RP_1(RP_{11}\ RP_{12} \ldots RP_{1m})$，则句法成分串：$RP_{11}\ RP_{12} \ldots RP_{1m}$

---

* 交段成分在交段结构中的位置是自左向右从 0 开始计数的，v —— 动词，n —— 名词.

$RP_2 \ldots RP_n$就组成了一个可能的交段左向组合实例(如图1(a)所示),其中$RP_{1n}$为交段成分.同理,如果$RP_n$是一个非叶子节点,且其子树结构为$RP_n(RP_{n1}\ RP_{n2} \ldots RP_{nk})$,则句法成分串:$RP_1\ RP_2 \ldots RP_{n-1}\ RP_{n1}\ RP_{n2} \ldots RP_{nk}$就组成了一个可能的交段右向组合实例(如图1(c)所示),其中$RP_{n1}$为交段成分.

### 2.2 松散结构SPR的发现

考虑分析树中的子树结构$PH(RP_1\ RP_2 \ldots RP_n)$,其中$PH$为子树根节点,$RP_i, i \in [1,n]$为其子节点成分.在分析树中搜索得到$PH$所有的左邻接成分$A_i$,如果$(A_i\ PH)$不是分析树中某个节点成分的前缀,即不存在这样的子树结构$X(A_i\ PH \ldots)$,则$A_i\ RP_1\ RP_2 \ldots RP_n$就组成了一个可能的交段右向组合实例(如图1(d)所示),其中$RP_1$为交段成分;同理,在分析树中搜索得到$PH$所有的右邻接成分$C_i$,如果$(PH\ C_i)$不是分析树中某个节点成分的后缀,即不存在这样的子树结构$X(\ldots PH\ C_i)$,则$RP_1\ RP_2 \ldots RP_n C_i$就组成了一个可能的交段左向组合实例(如图1(b)所示),其中$RP_n$为交段成分.

## 3 SPR的自动获取算法

上一节介绍的SPR发现算法需要利用树库语料,而大规模树库的构造往往需要花费大量的人力和物力,更何况现在还没有一个较大规模的汉语树库.因此,进行从原始文本(即经过了切分和词性标注处理的汉语文本)中自动获取SPR信息的研究,就显得很有必要了.自动获取SPR的关键是如何在原始句子上快速构造出合理的句法结构树(或森林)以及如何准确地计算不同歧义结构在分析树(森林)中的预期分布频度.对此,我们进行了以下处理.

### 3.1 原始文本的预处理

在文献[6,7]中,作者曾提出一种基于统计的汉语自动句法分析方法,它通过成分边界预测、括号匹配和分析树排歧等阶段的处理,可以自动生成原始句子的最佳分析树.在此基础上,文献[8]进一步提出可以利用一种改进的Inside-Outside算法在原始文本基础上自动训练得到汉语的PCFG(probabilistic context-free grammar)规则.利用这些研究成果,我们构造了一个原始文本预处理工具,其主要特点是可以自动分析产生原始句子的完整分析树(森林),它以Tomita的压缩共享森林(PSF)[9]形式表示出来,而不仅仅是选取其中的一棵最佳分析树,并且分析树(森林)中的每个成分节点上都带有成分内概率和外概率(有关内概率、外概率以及后面的成分预期分布频度的详细定义和计算方法可参阅文献[8])信息,其计算利用了汉语PCFG规则的概率信息.

### 3.2 交段结构的分布频度计算

对正确的分析树来说,其中交段结构的分布频度是确定的.而对PSF,情况就不同了,因为我们不知道其中哪棵树是正确的分析树.在这种情况下,唯一可利用的信息就是一个句法成分$A \rightarrow \lambda$在PSF中的预期分布频度$E(A \rightarrow \lambda)$.据此,可以按照如下方法计算交段结构的预期分布频度.

(1) 紧合结构SPR中的交段结构.其分布频度是由主成分和子成分的预期分布频度共同确定的.例如,对交段左向组合实例(如图1(a)所示),有:$E(\{PreIC\ IC\ SufIC\}) = E(A \rightarrow PreIC\ IC) \cdot E(B \rightarrow A\ SufIC)$.

(2) 松散结构SPR中的交段结构.其分布频度是由交段组合成分和其相邻左(右)成分的预期分布频度共同确定的.例如,对交段左向组合实例(如图1(b)所示),有:$E(\{PreIC\ IC\ SufIC\}) = E(A \rightarrow PreIC\ IC) \cdot E(SufIC)$,其中$E(SufIC) = \sum E(SufIC \rightarrow \lambda)$.

## 4 实验结果分析

本实验采用了以下语料:(1) 汉英机器翻译研究的测试题库,规模为1 434个汉语句子,约11 821个词,汉字总数为17 058,平均句长为8.243词/句.(2) 新加坡小学语文课本语料**,总规模为4 139个句子,约52 609个词,汉字总数为72 434,平均句长为12.711词/句.

选用这两部分语料的主要目的是因为它们都已标上了正确的分析树结构[6],即为树库语料,从而在此基础

** 此语料的电子版本由国立新加坡大学赖金定博士提供,在此表示感谢.

上可以方便地进行两种不同方式的 SPR 获取实验:分析树的 SPR 发现实验(实验 1)和原始文本的 SPR 自动获取实验(实验 2).通过对两者实验结果的比较,可以对目前的 SPR 自动获取算法的可行性和实用性进行比较深入的分析和讨论.

### 4.1 SPR 获取实验

在所有 5 573 句的语料文本上,进行两种 SPR 获取实验,得到以下结果.实验 1 共发现 4 021 个不同的 SPR 项,它们形成 SPR 表 A;实验 2 则发现了 6 091 个,通过设置选择阈值 $\mu=0.25$,将总频度(TF)(SPR 项的总频度(TF)定义为它的交段左向组合频度(LF)和右向组合频度(RF)之和,即 $TF=LF+PF$)小于 $\mu$ 的 SPR 项排除掉,得到 3 832 个,它们形成 SPR 表 B.

根据其 $TF$ 值的不同,可将 SPR 表 A 分成 3 个子表:(1) 高频 SPR:$TF>10$;(2) 中频 SPR:$1<TF\leqslant 10$;(3) 低频 SPR:$TF\leqslant 1$.然后通过检查子表中的每个 SPR 在 SPR 表 B 中是否出现,并分别记录其出现的总项数,可以计算出 3 个子表的 SPR 自动获取召回率.表 1 是具体的实验结果.从中可以看出,对于树库中出现的高频 SPR,98%以上都可以由自动算法所获取,对于中频 SPR,其召回率也达到了 85%以上,而对于低频 SPR,虽然其召回率只有 61%左右($\mu=0.25$),但这与实际的语言事实也是相符合的,因为这些 SPR 大多是一些特例结构,因此,在 PSF 中的预期频度必然是很小的,这样,其中的大部分就会由于阈值 $\mu$ 的设置而被排除掉.但即使在 $\mu=0.25$ 的条件下,SPR 自动获取算法的整体召回率也达到了 84%,这表明目前的自动算法所获取的 SPR 信息总体上是可信的.

表 1 SPR 表 A 的不同子表的自动获取召回率

| | $TF>10$ | $1<TF\leqslant 10$ | $TF\leqslant 1$ | 合计 |
|---|---|---|---|---|
| 表 A 的 SPR 项数 | 785 | 1 852 | 1 384 | 4 021 |
| 表 B($\mu=0.25$)中出现的 SPR 项数 | 773 | 1 591 | 840 | 3 204 |
| SPR 召回率($\mu=0.25$) | 0.984 7 | 0.859 1 | 0.606 9 | 0.836 1 |
| 表 B($\mu=0$)中出现的 SPR 项数 | 783 | 1 789 | 1 199 | 3 771 |
| SPR 召回率($\mu=0$) | 0.997 5 | 0.966 0 | 0.866 3 | 0.937 8 |

为了进一步分析表 A 和表 B 中不同的 SPR 特例的分布差异情况,我们从两个表中分别选择了总频度最高的前 10 个 SPR 项,并计算不同 SPR 项中的交段左向和右向组合概率 $LP$ 和 $RP$,得到表 2 的结果.从中可以看出,它们基本上是汉语中常见的交段歧义现象.尽管由于两种算法在频度计算方法上的不同:从最佳分析树中获取绝对频度和从 PSF 中获取预期频度,而造成一定的 $LF$ 和 $RF$ 差异,但其 $LP$ 和 $RP$ 的分布却是极为相似的:在两者共有的 8 个交段结构中,除了"v ^ v n"外,$LP$ 和 $RP$ 值都非常接近.这表明自动获取的数据很好地反映了树库语料中实际的 SPR 分布情况.

表 2 两个 SPR 表中总频度最高的前 10 个 SPR 项

| SPR 表 A | | | | | SPR 表 B | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 交段结构 | $LF$ | $RF$ | $LP$ | $RP$ | 交段结构 | $LF$ | $RF$ | $LP$ | $RP$ |
| d ^ vp wD vp | 487.00 | 3.00 | 0.99 | 0.01 | m ^ q n | 469.07 | 0.00 | 1.00 | 0.00 |
| m ^ q n | 476.00 | 0.00 | 1.00 | 0.00 | d ^ v np | 0.08 | 365.96 | 0.00 | 1.00 |
| r ^ vp wD vp | 441.00 | 3.00 | 0.99 | 0.01 | d ^ vp wD vp | 361.61 | 4.00 | 0.99 | 0.01 |
| v ^ v n | 140.00 | 292.00 | 0.32 | 0.68 | r ^ vp wD vp | 337.54 | 5.94 | 0.98 | 0.02 |
| d ^ v vp | 24.00 | 377.00 | 0.06 | 0.94 | v ^ v n | 147.66 | 134.24 | 0.52 | 0.48 |
| v ^ v np | 203.00 | 182.00 | 0.53 | 0.47 | v ^ v np | 151.87 | 128.95 | 0.54 | 0.46 |
| d ^ v np | 3.00 | 343.00 | 0.01 | 0.99 | d ^ v n | 0.00 | 272.22 | 0.00 | 1.00 |
| pp ^ v v | 0.00 | 326.00 | 0.00 | 1.00 | r ^ q n | 272.29 | 0.00 | 1.00 | 0.00 |
| d ^ v n | 2.00 | 302.00 | 0.01 | 0.99 | d ^ v vp | 17.03 | 246.64 | 0.06 | 0.94 |
| d ^ v v | 3.00 | 300.00 | 0.01 | 0.99 | v ^ np wD vp | 243.19 | 0.03 | 1.00 | 0.00 |

(在以上的交段结构实例中,以'^'标识交段成分.其中有关的词类标记和句法标记简单说明如下:d——副词,wD——句间停顿标点,m——数词,q——量词,n——名词,v——动词,r——代词;vp——动词短语,np——名词短语,pp——介词短语.具体内容可参阅文献[6].)

从以上的分析可以看出,从原始文本中自动获取的SPR信息,无论是从其总体的召回率数据,还是从其局部的高频交段结构的分布来看,都很好地反映了树库中实际的SPR分布特点.这表明目前的SPR自动获取算法具有较好的可行性和实用性.

### 4.2 SPR获取的收敛性

本节的实验试图回答这样的问题:从语料中获取的不同SPR数目是否随着语料规模的不断扩大而逐步收敛?需要提供多少训练语料才能达到收敛?

通过对目前的实验语料的均匀抽样,我们形成了11个测试样本,平均每个样本包含约507个汉语句子.从第1个样本开始,每次增加1个样本,记录在不同的样本条件下获取的不同SPR的项数,形成了图2的SPR项数随样本容量不断扩大的变化曲线.从图中可以看出,随着训练语料规模的不断扩大,发现的SPR项数也在不断增大,但其增长速度在不断减慢,显示出逐步收敛的趋势.尽管由于目前训练语料规模的限制,此趋势并不是很明显.

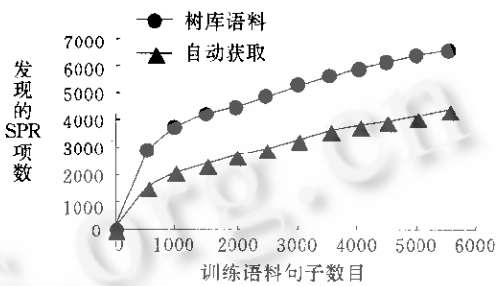


图2　SPR项数与测试语料规模的关系

## 5 结　语

本文提出了结构优先关系(SPR)的基本概念,它能对汉语歧义结构及其分布进行定量描述,并可通过两种不同的算法从树库语料和原始文本中自动获取.目前的实验结果显示,它在歧义结构研究方面具有一定的实用性.利用SPR信息,可以自动发现典型歧义结构,从大规模语料库中自动提取不同歧义结构的组合实例以及构造基于局部优先的优化分析器等.这些都显示出它在语言学研究和自然语言处理方面所具有的广阔的应用前景.

当然,目前的研究还只是初步的,没有涉及到对句法组合关系歧义(即同一结构可以组合为不同句法成分的歧义现象,如汉语中的"v+n"结构既可组成定中结构的np短语:工作时间,也可组成述宾结构的vp短语:看电影)、语义组合层次歧义和语义组合关系歧义[5]的自动获取问题.另外,如何对自动获取的SPR进行更深入的语言学分析,如:其中哪些是"潜在歧义结构"[10]?哪些是真歧义、伪歧义和准歧义结构[11]?都有待进一步的探索.

**致谢**　本文的研究得到国家自然科学基金资助,此项目编号为69705005.

### 参考文献

1 Hindle D, Rooth M. Structural ambiguity and lexical relations. Computational Linguistics, 1993,19(1):103～120

2 Basili R, Pazienza M T, Velardi P. Semi-automatic extraction of linguistic information for syntactic disambiguation. Applied Artificial Intelligence, 1993,7:339～364

3 Collins M, Brooks J. Prepositional phrase attachment through a backed-off model. In: Yarowsky D, Church K eds. Proceedings of the 3rd Workshop on Very Large Corpora. Cambridge, Massachusetts: Massachusetts Insitute of Technology, 1995. 27～38

4 朱德熙.汉语句法中的歧义现象.中国语文,1980,(2):81～92
(Zhu De-xi. Ambiguous phenomena of Chinese syntax. Chinese Language, 1980,(2):81～90)

5 黄国营.现代汉语的歧义短语.语言研究,1985,(1):69～89
(Huang Guo-ying. Ambiguous phrases in contemporary Chinese. Language Research, 1985,(1):69～89)

6 周强.汉语语料库的短语自动划分和标注研究[博士学位论文].北京大学,1996
(Zhou Qiang. Phrase bracketing and annotating on Chinese language corpus[Ph. D. Thesis]. Beijing University, 1996)

7 Zhou Qiang. A statistics-based Chinese parser. In: Zhou Joe, Church K eds. Proceedings of the 5th Workshop on Very Large Corpora. Tsinghua University, Beijing, August 1997. 4～15

8 周强,黄昌宁.汉语概率型上下文无关语法的自动推导.技术报告TR97002,清华大学计算机系,1997

(Zhou Qiang, Huang Caang-ning. An inference approach for Chinese probabilistic context-free grammar. Technical Report TR97002, Department of Computer Science, Tsinghua University, 1997)

9 Tomita M. Efficient Parsing for Natural Language — a Fast Algorithm for Practical System. Boston: Kluwer Academic Publishers, 1986

10 冯志伟. 中文科技术语的结构描述及潜在歧义. 中文信息学报, 1989, 3(2): 1～15
(Feng Zhi-wei. Structural description of Chinese scientific terms and potential. Journal of Chinese Information Processing, 1989, 3(2): 1～15)

11 詹卫东. 现代汉语 VP 的结构定界和结构关系判定[硕士学位论文]. 北京大学, 1996
(Zhan Wei-dong. Determining boundaries and constructional relations of verb phrase in contemporary Chinese[M. S. Thesis]. Beijing University, 1996)

## Automatic Acquisition for Chinese Structure Preference Relations

ZHOU Qiang HUANG Chang-ning

(*Department of Computer Science and Technology Tsinghua University Beijing* 100084)
(*State Key Laboratory of Intelligent Technology and Systems Tsinghua University Beijing* 100084)

**Abstract** A new ambiguity representation scheme SPR (structure preference relation) is proposed in this paper, which consists of useful quantitative distribution information for ambiguous structures. Two automatic acquisition algorithms, (1) acquired from treebank, (2) acquired from raw texts, are introduced, and some experimental results which prove the availability of the algorithms are also given. At last, some SPR applications in linguistics and natural language processing are introduced and some future research directions are proposed in this paper.

**Key words** Structure preference relation, ambiguous structure, knowledge acquisition, corpus.