句法标注的一般模型与参数分析*)

李良炎　何中市
(重庆大学语言认知及信息处理研究所　重庆 400044)

摘　要　句法标注是语料标注的重点、难点所在，必须以一定的句法理论为基础。短语结构语法和依存语法是句法标注的基础理论，彼此却有很大的不同。本文以形式化为目标，提出句法标注的一般模型，分析比较两种句法标注的参数异同，深刻揭示了基于短语结构语法和基于依存语法的句法标注与一般模型之间的关系，并提出阅读依存中心原则，力图解决基于依存语法的句法标注难以标注缺省结构的问题。

关键词　句法标注，短语结构语法，依存语法，阅读依存中心原则

General Model with Parameters Analysis of Syntax Tagging

LI Liang-Yan　HE Zhong-Shi
(Institute for Language, Cognitive & Information Processing, Chongqing University, Chongqing 400044)

Abstract　Syntax tagging is the main and most difficult point of corpus tagging, and should be based on syntax theory. Phrase Structure Grammar (PSG) and Dependency Grammar (DG) are the basic theories of syntax tagging, but they vary in many important ways. Aiming at formalization, this paper presents a general model of syntax tagging, and analyzes the parameters of the syntax tagging models based on PSG and DG. Meanwhile, Reading Dependency Head Principle (RDHP) is proposed to try to improve on the syntax tagging model based on DG.

Keywords　Syntax tagging, Phrase structure grammar, Dependency grammar, Reading dependency head principle

1　引言

人工智能的瓶颈在知识，自然语言处理的瓶颈在语言知识。归根结底，语言知识来自人头脑中或现实媒介中的语言材料，即语料。在信息化时代，大量语言材料的获取、存储与加工成为可能，在自然语言处理与理解、人机对话、机器翻译、语音识别、语言学、语言教学、词典编撰、语言翻译等诸多研究或应用领域越来越受到重视，一门新兴学科——语料库语言学应运而生。

语言的理解有赖于人的经验。一般认为，以文本或语音形式存在、经过知识标注的语料可以借助计算机通过机器学习、数据挖掘等技术提取和分析更加丰富的知识。因此，有必要通过人工将知识表示出来，并标记到语料中的对应位置，这就是语料标注。语言可以分为字、词、句、段、篇等层次，不同层次上的标注既有联系又有区别。句法标注(Syntax Tagging)是对句子的语法结构进行标注，相对词的标注(Lexical Tagging)难度更大。

句法标注必须以语法理论为基础。一般认为，语言学研究中语法研究相对成熟，语义研究正在攻关，语用研究相对薄弱。然而，句法标注的现状和水平与这种看法是不相称的。根据目前语料库建设的进展来看，无论是标注规范、规模和质量，句法标注都远远落后于词的标注，难以满足研究和应用的需要。问题出在哪里？是语法理论本身有问题？还是语料库语言学自身的原因？语料库语言学是一门新兴学科，还没有成熟的理论，很容易受到相邻优势学科(如语言学)促进和制约的双重影响。因此，从语料库语言学的角度反思现有理论，对语料库语言学和相邻学科的发展都是有益的。

短语结构语法(Phrase Structure Grammar, PSG)和依存语法(Dependency Grammar, DG)是句法标注的两种基础理论，并形成了语法研究和句法标注的两大流派。基于PSG和DG句法标注是否可以概括到一般模型？根本的差异是什么？共同的因素和局限在哪里？继承和发展这两种句法标注理论的方向在哪里？本文试图从语料库语言学角度，以形式化的一般模型为基础，结合认知语法的研究成果回答这一系列问题。

2　句法标注的一般模型与参数

句法标注是在词标注之后进行的。词标注包括词的切分(汉语)、词类标注。词的切分相对句法标注的独立性较强，词类标注则与句法标注的关系十分密切。不同的句法标注体系往往有不同的词类标注体系。“目前许多句法赋码系统以词类赋码系统的输出为输入，……这在一定程度上隔离了词汇和句法的关系”[1](赋码即标注——笔者注)，基于此，本文将词标注狭义地定义为词的切分，而将词类标注纳入句法标注整体考虑。由于词类标注既可以是词性标注，也可以是词义标注，因此本文中的句法标注是广义的，即标注句子的语法结构或语义结构。

撇开具体的语法理论，本文提出句法标注的一般模型可形式化表示为：

输入：词串 $W=\{w_0, w_1, \cdots, w_m\}$，概念集 $C=\{c_0, c_1, \cdots, c_n\}$，知识库 $K=\{k_0, k_1, \cdots, k_l\}=R\cup E$，中心原则 $R_{HEAD}\in R$。R 为规则库，E 为实例库。

标注：以人工方式，或计算机自动标注辅以人工校正方式，根据 W、C、K 构造句法结构网络 G。计算机自动标注即

*)本文受国家自然科学基金项目(60173060)资助。**李良炎**　副教授，博士，主要研究方向：计算语言学、语料库语言学、认知语言学、艺术心理学；**何中市**　教授，博士，主要研究方向：自然语言处理、机器学习、概率论、容错计算。

句法自动分析(Syntax Analysis, SA),可采取基于 R 的规则技术(Rule Based, RB)、基于 E 的统计技术(Statistic Based, SB)、基于 E 的实例技术(Example Based, EB)、综合多种技术(Comprehensive Technique, CT)来实现。

输出:句法结构网络 $G=\{V, D, C^V, C^W, C^D\}$。结点集 $V=\{v_0, v_1, \cdots, v_i\}=W\cup C^V$,边集 $D=\{d_0, d_1, \cdots, d_j\}$,词类标注 $C^W=\{c_0^w, c_1^w, \cdots, c_m^w\}$,关系标注 $C^D=\{c_0^d, c_1^d, \cdots, c_j^d\}$,$C^V\in C, C^W\in C, C^D\in C$。

词串 W 是词的切分输出。概念集 C 是根据特定的语法理论预设的一套概念符号体系。G 的结点集 V 包括词汇 W 和概念 C^V,C^V 是用作顶点的概念;G 的边集 D 包括由顶点 V 构成的关系$\{\langle v1_{d0}, v2_{d0}\rangle, \cdots, \langle v1_{dj}, v2_{dj}\rangle\}$;$C^W$ 是所有词 W 的词类(概念)标注;C^D 是所有边的关系(概念)标注;C^V、C^W、C^D 均来自概念集 C。

知识库 K 是用以限定词汇 W 或概念 C 的关系的知识,包括根据特定的语法理论预设的规则库 R 和已标注实例库 E,是句法标注的依据。R_{HEAD} 为中心原则,规定句法结构的层次关系。例如,DG 根据谓语中心原则规定,主语、宾语、状语等成分从属于谓语成分。

计算机自动标注本质上就是句法自动分析,目前已有相对成熟的技术。基于 R 的规则技术(RB)可以采取自顶向下或自底向上的自动分析方法。基于 E 的统计技术(SB)需要事先以 E 为训练集计算统计模型。基于 E 的实例技术(EB)需要计算实例的相似度。

一个具体的句法标注模型决定于参数设定。本文提出的句法标注一般模型的参数表述如下:

(1)词串 W:随句法标注任务而变化。

(2)概念集 C:根据一定的语法理论设定。C 的元素个数为 $|C|$,是知识表示粒度的一个量化指标。相对来说,$|C|$ 越大,知识表示粒度越精细。

(3)知识库 K、R、E:从规模上讲,K 的元素个数为 $|K|$,是知识库完备性一个量化指标。相对来说,$|K|$ 越大,知识库越完备。从构成上讲,当 E 为空时 K 为纯规则库,当 R 为空时 K 为纯实例库,当 E、R 均不为空时 K 为混合库。

(4)句法分析 SA:可以是规则技术 RB、统计技术 SB、实例技术 EB、综合技术 CT。

(5)概念结点 C^V:当 C^V 为空,结点 V 当中只包括词汇 W,不包括概念 C^V。当 C^V 不为空,结点 V 当中既包括词汇 W,又包括概念 C^V。

(6)词类标注 C^W:当 C^W 不为空,必须标注词类,否则不用标注词类。当然,词类标注是语料标注不可缺少的一环,当 C^W 为空时,表明以另外的方式进行处理。

(7)关系标注 C^D:当 C^D 不为空,必须标注 G 的边 D 所表示的关系,否则不用标注。与词类标注类似,关系标注也是语料标注不可缺少的一环,当 C^D 为空时,表明以另外的方式进行处理。

(8)边集 D:D 决定 G 的网络拓扑结构。由 V、D 构成的网络通常是树结构,因为这种结构具有层次性,比较好地刻画了语言的层次性特征。目前的句法标注语料库多为树结构,称为树库(Tree Bank)[2~4]。

(9)中心原则 R_{HEAD}:R_{HEAD} 决定了结点 V 的相对层次关系。在 D 所决定的树结构中,结点 v_x 位于 v_y 的层次之上或之下不是随意的,必须根据中心原则 R_{HEAD} 唯一判定。R_{HEAD} 由特定的语法理论给出,并作为 R 的必备元素,用以指导句法结构 G 的构造。如果 D 为层次关系,那么 R_{HEAD} 是不可缺少的规则。

以上给出了句法标注的一般模型和参数,作为一个统一的标准来分析比较不同句法标注体系的异同,具有重要价值。通过分析,现有的主要句法标注体系都满足这一模型,有些参数是相同的,有些参数则不同。正是不同的参数导致不同句法标注系统具有迥异的特点。

纵观目前语料库标注项目,有两种占据主流的句法标注体系,基于短语结构语法的句法标注(PSG-based Syntax Tagging, PSGTG)和基于依存语法的句法标注(DG-based Syntax Tagging, DGTG)。以下分别对两种体系的模型参数进行比较分析。

表 1 PSGTG 与 DGTG 的参数异同

TG	C	K	R	E	SA	C^V	C^W	C^D	D	R_{HEAD}
PSGTG	小	任选	大	任选	任选	非空	空	空	层次树	左件中心
DGTG	大	任选	小	任选	任选	空	非空	非空	层次树	补足中心

3 基于短语结构语法的句法标注(PSGTG)

美国语言学家乔姆斯基 (Noam Chomsky)于 1957 年出版专著《句法结构》,从而奠定了短语结构语法(PSG)的理论基础。其后发展起来的许多语法理论可以直接或间接归到这一流派,如中心词驱动的短语结构语法(HPSG)、广义短语结构语法(GPSG)等。到目前为止,PSG 仍然是最重要的句法结构标注理论,为世界上众多语料库项目所采用和发展。

以本文提出的句法标注一般模型为参照来看,PSGTG 可以用这一模型进行解释,其模型参数特征见表 1。

PSGTG 的主要参数特征是:

(1)C 相对较小,R 相对较大:重视规则库 R 的建设,作为语言知识的主要载体。知识表示粒度 C 不能太大,否则 R 就会出现"规则爆炸",规则的描写和维护变得困难(这是规则技术的瓶颈问题)。C 一般为粗粒度的语法概念,例如名词 N、动词 V、名词短语 NP、动词短语 VP 等。PSGTG 的 C、R 不太适合描写细粒度的语义概念和语义规则。

(2)C^V 非空:句法结构的树结点包括词汇以外的概念,如名词短语 NP、动词短语 VP 等。这使得 PSGTG 的结点数相对 DGTG 要多。

(3)C^W、C^D 为空:这是由于 PSGTG 的词类标注和关系标注均以重写规则的形式表示在 R 中。

(4)R_{HEAD} 为左件中心原则:PSG 的重写规则形式为 $A\rightarrow B$。左件中心原则规定,在句法结构 G 中,A 是 B 的父结点。根据左件中心原则构造的句法结构体现了句子各成分的纵向层次关系(句子由短语构成、短语由词构成),符合人们的普遍语感。

总的来看,PSGTG 非常强调规则,这是 PSG 由形式语法发展而来,强调语言普遍规则所决定的。但在具体的语料库句

法标注项目中，规则技术、统计技术、实例技术其实并不矛盾。实践表明，反而可以取长补短，综合提高。表1中E、SA都是任选的，这其实是PSGTG的发展方向。因此，C和R的大小都是相对的，并不是综合技术意义上的PSGTG的关键特征。相对DGTG而言，PSGTG的关键特征是(2)、(3)、(4)。

以句子“我看书”为例，其PSGTG可描述为：

输入：词串 $W=\{$我，看，书$\}$，概念集 $C=\{$S|句子，N|名词，NP|名词短语，V|动词，VP|动词短语，+|语言及概念组合操作，→|重写规则操作$\}$，重写规则知识库 $K=R=\{R_{HEAD}$|左件中心原则，$S\rightarrow NP+VP$，$VP\rightarrow V+NP$，$NP\rightarrow N$，$VP\rightarrow V$，$N\rightarrow$我，$V\rightarrow$看，$N\rightarrow$书$\}$。重写规则操作“→”，指用右边的表达式代替左边的表达式得到新的表达式，$A\rightarrow B$称为重写规则。

处理：主要通过计算机自动分析(自底向上、自顶向下等规则分析技术[5])辅助人工校正方式，根据 W、C、K 构造 G。

输出：句法结构 $G=\{V, D, C^V, C^W, C^D\}$，$V=\{$我，看，书$\}\cup C^V$，$D=\{d_0, d_1, \cdots, d_8\}$，$C^V=\{N1, N2, NP1, NP2, V, VP, S\}$，$C^W$为空，$C^D$为空，如图1所示。

不难看出，词类标注信息存储在边 d_5、d_6、d_8 当中，因此 C^W 为空；关系标注信息存储在结点 S(对应 $S\rightarrow NP+VP$)、VP(对应 $VP\rightarrow V+NP$)中，因此 C^D 为空。根据左件中心原则 R_{HEAD}，重写规则的左件总是右件的父结点。这些体现了PSGTG独特的知识表示和标注策略。

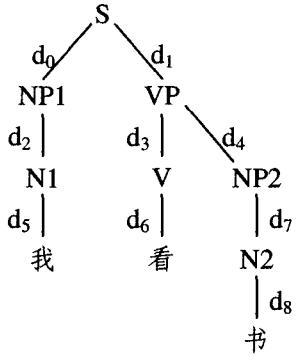

图1 PSGTG实例

4 基于依存语法的句法标注(DGTG)

法国语言学家特思尼耶尔(Lucien Tesnière)于1959年出版专著《结构句法基础》，从而奠定了依存语法(DG)的理论基础。其后发展起来的许多语法理论可以直接或间接归到这一流派，如词汇依存语法(WD)、概念依存理论(CD)、核心依存理论(KD)[6]等。由于相对PSG而言，DG更简洁、直观、经济，适应性更强，因此反而有后来居上之势，目前已经成为世界上较为通用的句法标注理论。

以本文提出的句法标注一般模型为参照来看，DGTG可以用这一模型进行解释，其模型参数特征见表1。

DGTG的主要参数特征是：

(1)C相对较大，R相对较小：重视概念集 C 的定义，作为语言知识的主要载体。DGTG对依存关系 C^D 进行详尽分类和定义就是一种表现[7]。R 概括性很高，主要以公理的形式限定句法结构层次关系。由于这种处理方式，DGTG不存在PSGTG的“规则爆炸”问题，知识表示粒度可以很精细，适合描写细粒度的语义概念和语义规则。由于语义研究已经成为语言学、计算语言学、语料库语言学等领域研究的热点，因此DG越来越受到重视。

(2)C^V 为空：句法结构的树结点不包括词汇以外的概念，是DGTG最明显的特点。这使得DGTG的结点数相对PSGTG要少得多。

(3)C^W、C^D 不为空：DGTG的词类标注和关系标注并不像PSGTG那样以结点 V 或边 D 的形式表示出来，而是作为 V 或 D 的属性，属于复杂特征集模式。

(4)R_{HEAD}为补足中心原则：补足中心原则规定，在句法结构 G 中，词 A 如果是词 B 的补足成分，那么 B 是 A 的父结点。谓语中心原则是最重要的补足中心原则，如果词 A 是主语、宾语、状语等补足成分，词 B 是谓语，那么 B 是 A 的父结点。根据补足中心原则构造的句法结构体现了句子各成分的横向组合关系(词与词通过语法或语义上的补足关系构成短语和句子)，符合人们的普遍语感。

(5)R_{AXIOM}为公理系统：$R_{AXIOM}\in R$，用以限定句法结构一般结点关系。19世纪70年代美国计算语言学家Robinson提出了DG的四条公理[8]。公理1：一个句子中只有一个成分是独立的；公理2：除独立成分外，句子中其他成分都必须依存于某成分；公理3：句中任何一个成分都不能依存两个以上的其他成分；公理4：如果A成分从属于B成分，而C成分在句中处于A和B之间，则C成分或者从属于A，或者从属于B，或者从属于A，B之间的某个成分。荷兰机器翻译专家K. Schubert于1987在Robinson的基础上提出了12条公理[8]。R_{AXIOM}是DGTG为数不多但必不可少的规则。

总的来看，DGTG非常强调概念，这是DG注重语言描写，强调语言实际经验所决定的。但规则技术、统计技术、实例技术并不矛盾，可以取长补短，综合提高。R_{HEAD}支持人工标注，却难以和 R_{AXIOM} 一样支持计算机自动分析。DGTG如果要实现基于 RB 的计算机自动标注，就有必要也有可能建设相对完备的规则库 R。因此，C 和 R 的大小都是相对的，并不是综合技术意义上的DGTG的关键特征。相对PSGTG而言，DGTG的关键特征是(2)、(3)、(4)、(5)。

以句子“我看书”为例，其DGTG可描述为：

输入：词串 $W=\{$我，看，书$\}$，概念集 $C=\{v|$动词，$n|$名词，$SBV|$主谓关系，$VOB|$动宾关系，+|语言及范畴组合操作，/|属性编码操作$\}$，知识库 $K=R=\{R_{HEAD}|$谓语中心原则，$R_{AXIOM}|$公理系统，$(n+v)/SBV$，$(v+n)/VOB$，我$/n$，看$/v$，书$/n\}$。属性编码操作“/”，即给特定语言表达式作上特定概念的属性标记，A/C 称为编码规则。

处理：通过人工方式根据 W、C、R 构造 G。如果给出知识库 K 中的实例库 E 不为空，或者规则库 R 比较完备，也可以采取计算机自动分析[9, 10]辅以人工校正的方式。

输出：句法结构 $G=\{V, D, C^V, C^W, C^D\}$，其中 $V=\{$我，看，书$\}$，C^V 为空，$D=\{d_0, d_1\}$，$C^W=\{n_0, v, n_1\}$，$C^D=\{SBV, VOB\}$，如图2所示。

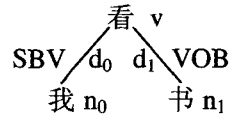

图2 DGBTG实例

对比图1不难看出，结点集 V 为词 W，词类标注 C^W 为结点集 V 的属性，关系标注 C^D 为边集 D 的属性，结点关系和层

次符合 R_{HEAD} 和 R_{AXIOM} 原则。这些体现了 DGTG 独特的知识表示和标注策略。

5 引入阅读依存中心原则的句法标注(RDGTG)

以上参数分析表明,对差异很大的 PSGTG 和 DGTG 可以用句法标注的一般模型来描述,只是一些参数侧重点和取值有所区别。

总的来看,DGTG 的优越性表现在两个方面:

(1)比 PSGTG 更简洁、直观、经济:比较图 1、图 2 可知。为了确保句法标注质量,在大规模语料库加工过程中尤其早期建立训练集(即实例库 E)时人工方式必不可少。因此,一种存储空间小,人工操作简便直观的句法标注模型显然更受欢迎。这是 DGTG 有后来居上之势的主要原因。

(2)比 PSGTG 适应性更强:DGTG 重视概念集 C 的定义,降低了 R 的负担,避免了"规则爆炸"风险,符合"小规则、大词库"的原则。既可以描述语法,也适合描述语义,具有适应性。这是目前进行语义标注的语料库一般采取 DGTG 的主要原因。

不过,在具体的句法标注实践中 DGTG 还是暴露出一些问题,"对一些没有明确依存关系的成分,标注起来则有些力不从心"[11],存在"依存失败"[7]现象,最突出的是难以标注缺省结构。缺省结构标注是任何句法标注都会面临而且很难解决的问题。

人类的语言符合经济性原则,而缺省结构恰恰体现了这一原则。借助上下文省略一些成分,人们仍然能够理解,但对计算机来说却是一种挑战。句法标注的根本目的是让计算机能够正确提取语料的语法和语义信息。缺省结构在真实语料中大量出现,句法标注是不能回避的。在很多情况下,DGTG 不但不能正确标注缺省结构,反而在中心原则 R_{HEAD} 和公理 R_{AXIOM} 的强制限定下给出违背真实语法或语义结构的标注结果,形成干扰信息。请看以下 4 个句子:

句 1　我看一下书

句 2　(真是好书啊?)我看一下

句 3　我看一本书

句 4　(好多书啊!)我看一本

句 2 是句 1 宾语省略句,句 4 是句 3 宾语省略句。各句的 DGTG 见图 3(为简便起见,把"一下"、"一本"作为一个词处理,省略了词类标注 C^W 和关系标注 C^D):

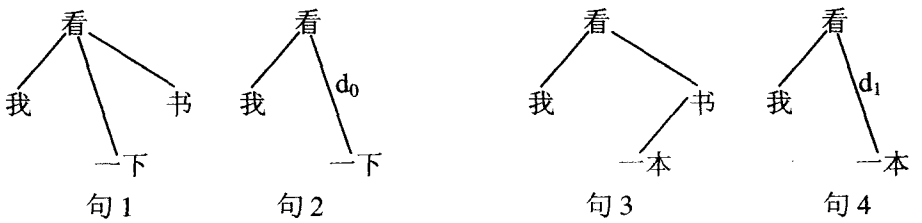

图 3　句 1～4 的 DGTG

问题出在句 4。句 1 和句 3 的依存结构是不同的,然而句 2 和句 4 却有了相同的依存结构。因为句 4 省略了"书",根据 DGTG 的 R_{AXIOM} 公理 1、2,"一本"必须依存于独立谓语成分"看"。于是"看一本"和"看一下"依存结构相同,实际上违反了句 3 的正确结构。当然,我们可以采取补救措施,为 d_1 标注一个特殊的依存关系属性 C^{error}(即依存失败),但这显然不是好办法。

美国认知语言学家兰盖克(Ronald W. Langacher)分别于 1987 年、1991 年出版专著《认知语法基础》一、二卷,开创了认知语法理论,关于语法结构有如下观点:

"如果一个构件 A 使另一构件 B 的一部分抽象变为具体,那么构件 A 就叫做概念自主(Conceptually Autonomos)的构件,构件 B 就叫做概念依存(Conceptually Dependent)的构件"[12]。

举例来说:独立地看,"一本"隐含一个抽象的、可数的、可用"本"量化的事物,可表示为"一本(x)"。"书"使"x"变得具体,因此"书"是概念自主的,"一本"是概念依存的。

从信息表达的角度来看,"书"表达了相对完整而具体的信息,因此是概念自主的;"一本"表达了不完整不具体的信息,因此是概念依存的。

从数学表达式的角度来看,"一本"类似函数,"书"类似参数,函数的地位显然是第一位的,决定了对参数的处理过程和返回参数。例如,"旧书"与"一本书"的区别不在"书",而在"旧"和"一本"。再从阅读认知过程来看,当人们读到"一本"时,实际上已经在期待"一本"后面那个具体事物跟着出现。为什么我们觉得"我看一本"是缺省句?因为"看"和"一本"相对"书"都是概念依存的,因此人们会判定,"我看一本"的缺省成分可能是"书"。而读到"我看书"时,人们不会认为这是一个省略句,因为"书"表达的信息已经自足了。

正是从这两个角度来看"一本"与"书"的关系,有足够的理由认为在句法结构中,"一本"应是"书"的父结点,而不是按传统的补足中心规则,中心成分总是限定成分的父结点(如图 3 中句 3 所示)。依存成分是自主成分的父结点,本文将这一原则命名为阅读依存中心原则(Reading Dependency Head Principle, RDHP),并引入 DGTG 替换补足中心原则。以 RDHP 为参数的 DGTG 称为 RDGTG。

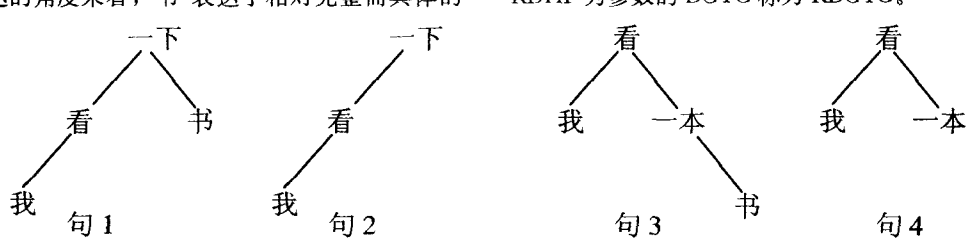

图 4　句 1～4 的 RDGTG

(下转第 238 页)

sign. MIT Press, 1991. 483～514
30 Smith D R. KIDS: a semi-automatic program development system. IEEE Trans on Software Engineering -Special Issue on Formal Methods, 1990, 16(9)
31 Wang T C, Goldberg A. A mechanical verifier for supporting the design of reliable reactive systems. In: Int'l Symp. on Software Reliability Engineering, Austin, TX, May 1991
32 Blaine L, Goldberg A. DTRE: a semi-automatic transformation system. In: Möller B, ed. Constructing Programs from Specifications. North Holland, May 1991
33 Kestrel Institute and Kestrel Development Corporation. Specware Language Manual. http:// www. specware. org/doc. html, Oct13,2006
34 Smith D R. Toward a classification approach to design, In: Proc. 5th Int'l Conf. on Algebraic Methodology and Software Technology, AMAST'96, LNCS 1101, Springer Verlag, 1996
35 Smith D R. Mechanizing the development of software, calculational system design. In: Broy M, ed. Proc. Int'l Summer School Marktoberdorf, NATO ASI Series, IOS Press, Amsterdam, 1999
36 Schmid U. Inductive Synthesis of Functional Programs. In: Carbonell J G, Siekmann J, eds. LNCS 2654, Springer-Verlag, 2003
37 Xue Jinyun. Two new strategies for developing loop invariants and their applications. Journal of Computer Science and Technology, 1993, 8(2): 147～154
38 Xue Jinyun. A unified approach for developing efficient algorithmic programs. Journal of Computer Science and Technology, 1997, 12(4): 314～329
39 Xue Jinyun. A practicable approach for formal development of algorithmic programs. In: Lu Jian, Noro M, eds. Proc. Int'l Symp. on Future Software Technology (ISFST'99), Software Engineers Associations of Japan, Oct. 1999
40 Xue Jinyun, Ruth D. A derivation and proof of Knuth's binary to decimal program. Software: concepts and tools, 1997, 18: 149～156
41 Xue Jinyun, Ruth D. A simple program whose derivation and proof is also. In: Proc1st IEEE Int'l Conf. on Formal Engineering Method (IEEE ICFEM'97), IEEE CS Press, 1997. 132～139
42 Xue Jinyun. Formal derivation of graph algorithmic programs using partition-and-recur. Journal of Computer Science and Technology, 1998, 13(6): 553～561
43 Meertens L. Algorithmics Towards programming as a mathematical activity. In: Mathematics and Computer Science. Proc. CWI Symp. November 1983
44 Paige R. Viewing a program transformation system at work. In: Hermenegido M, Penjam J, eds. Proc. 6th Int'l Symp. on Programming Language Implementation and Logic Programming, Springer Verlag, 1994. 5～24
45 Bellegarde F. ASTRE: towards a fully automated program transformation system. In: Hsiand Jieh, ed. Proc. 6th Int'l Conf. on Rewriting Techniques and Applications, LNCS 914, Springer Verlag, 1995. 403～407
46 http://www. cse. dmu. ac. uk/～mward/fermat. html, October 13, 2006
47 Lüth C, Kolyang T H, Brückner B K. TAS and IsaWin: tools for transformational program development and theorem proving. In: European Joint Conference on Theory and Practice of Software (ETAPS'99). Amsterdam, LNCS 1577, Springer, March 1999
48 http://www. program-transformation. org. October 2006

(上接第 192 页)

仅仅引入 RDHP 是不够的，句法标注一般模型中的其他参数可能也需要改变，这有待深入研究。例如，“看(x)”和“一本(x)”这两个表达式的返回参数是不同的，“一本(x)”返回参数为“x”，“看(x)”返回参数为“看”。正因为如此，表达式“看(一本(书))”成立，“一本(看(书))”不成立。“看”与“一本”的这种区别可以在 R 中定义，在 C^W 中加以标注。另外，表达式“(x)一下”的返回参数为“x”，即“看”；表达式“(x)看”的返回参数为“看”。根据这些知识定义，句 1、2、3、4 的 RDGTG 见图 4。

根据函数、输入参数、返回参数的关系，各句结构的逆构造过程如下：

句 1　我看一下书：(((我)看(x))一下)(书) = ((看(x))一下)(书) = 看(x)(书) = 看(x = 书)；

句 2　我看一下：((我)看(x))一下 = (看(x))一下 = 看(x)；

句 3　我看一本书：((我)看(x))(一本(书)) = 看(x)(书) = 看(x = 书)；

句 4　我看一本：(我)看(一本(x)) = 看(x)。

句 1 和句 3 的 x 有明确取值，为完整句。句 2 和句 4 则是缺省句。可见，基于看(x)和一本(x)的阅读依存知识，可以预测并判定缺省结构及其成分。

看起来，RDGTG 与 DGTG 的标注结果有了很大的差异，而且不符合补足中心原则，以及在这种原则影响下人们长期以来形成的语感。但这种结构符合人们阅读认知经验，而且可以按函数标准给出形式化地解释，其解释结果符合句子本身的语法和语义结构，没有错误和干扰信息。因此，RDGTG 更适合计算机处理，更符合句法标注的本来目的。

结束语　语料标注从根本上讲是为计算机处理服务的，因此必须坚持形式化。本文采取形式化方法，提出了句法标注的一般模型和参数，并对 PSGTG 和 DGTG 的参数进行了分析。发现 PSGTG 和 DGTG 是可以用句法标注一般模型加以描述的，只是某些参数不同。相对来说，DGTG 更简洁、直观、经济，适应性更强，但其补足中心原则仍然有缺陷，存在依存失败现象，难以标注缺省结构。本文受认知语法的概念自主与依存观点启发，引入函数形式，提出一种更符合人们阅读认知语感的阅读依存中心原则 RDHP，替换补足中心原则，并在缺省结构标注中得到初步验证。本文力图建立一种新的句法标注模型 RDGTG。今后将进一步研究和改进 RDGTG 的参数，并用于目前在建的语料库项目进行验证。

参 考 文 献

1 杨惠中. 语料库语言学导论. 上海：上海外语教育出版社，2002. 151
2 周强. 汉语句法树库标注体系. 中文信息学报，2004，18(4)：1～8
3 Marcus M, Santorini B, et al. Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics, 1993, 19(2): 313～330
4 台湾中央研究院语言所中文句结构树资料库. http://turing. iis. sinica. edu. tw /tree-search/
5 冯至伟. 基于短语结构语法的自动句法分析方法. 当代语言学，2000，2(2)：84～98
6 由丽萍，范开泰，刘开瑛. 汉语语义分析模型研究述评. 中文信息学报，2005，19(6)：57～63
7 尤昉，李涓子，王作英. 基于语义依存关系的汉语语料库的构建. 中文信息学报，2003，17(1)：46～53
8 刘海涛. 依存语法和机器翻译. 语言文字应用，1997，23(3)：89～93
9 Zhou Ming. A block based dependency parser for unrestricted Chinese text. In: The Second Chinese Language Processing Workshop, ACL2000. Hong Kong, 2000. 78～84
10 王建会，王雷，胡运发. 词语间依存关系的定量识别. 中文信息学报，2005，19(4)：31～38
11 周强. 汉语句法树库标注体系. 中文信息学报，2004，18(4)：1～8
12 齐振海，张辉. 导读. 见：Langacker R W. 认知语法基础(理论前提). 北京：北京大学出版社，2004. 9. 13