现代汉语述语动词机器词典的研究(二探)*

陈群秀
清华大学计算机系

摘要:机器词典的规模和质量是决定自然语言处理系统成败的关键,其中述语动词词典又是整个机器词典的关键。清华大学计算机系和中国人民大学语文所抓住这关键的关键,联合研制一部现代汉语述语动词机器词典。本文首先简单介绍现代语法理论—原则参数语法,然后着重介绍联合研究小组在原则参数语法指导下,用计算语典学方法,对现代汉语常用的 2300 个动词的 3000 个义项用论旨网格方式作详细的描述,意欲建立一部信息丰富、结构合理的机器可循词典,并探讨基于大规模语料的自动编辑出版词典方法。

Research on A Machine Tractable Dictionary of Contemporary Chinese Predicate Verbs: A Second Pass

Chen Qunxiu
Dept. Computer Science & Technology, Tsinghua University, Beijing

Abstract: The quality of a natural language processing system is determined by the scale and quality of its machine dictionary, which is in turn determined by the quality of the machine dictionary of predicate verbs. Concentrating on the most crucial part, the Computer Science Dept. of Tsinghua University and the Language and Character Institute of People s University of China have worked together and compiled a machine dictionary of contemporary Chinese predicate verbs. The paper made a brief introduction to a modern grammar—Principles and Parameters Approach ——before it portrayed the work done by the combined research group. Guided by Principles and Parameters Approach and Computational Lexicography, the group has made a detailed description of 3,000 senses of 2,300 commonly used Chinese verbs using thematic—grid method, so as to develop a well structured machine tractable verb dictionary that will contain large amount of information. Moreover, the group has explored ways of automatic edition and publication of dictionaries on the basis of large—scale corpors.

一、导言

汉语是我国乃至世界范围的主要交际工具和信息载体,也是我国几千年文化和知识世代相传的主要媒体。使用计算机来进行汉语信息处理是我国进入信息社会的需要,也是衡量

* 国家自然科学基金资助项目,项目号 69383007。

我国科技实力的一个重要标志。信息高速公路(NII)的提出和建立,在全世界引起强烈的反响并将对人类的经济、生产、社会生活带来重大影响。而自然语言处理和理解的研究将是NII要解决的一个关键和瓶颈问题。

在计算语言学界,起来越多的专家把机器词典的规模和质量看作是决定一个自然语言处理系统成败的关键,其中述语动词 机器词典又是整个机器词典的关键,因为述语动词 是句子的核心,别的成分都跟它挂钩,被它吸收。因此建立一部信息丰富、结构合理的现代汉语述语动词 机器词典不仅是一项重要的基础理论研究,也是一项大规模的知识工程[1]。清华大学和中国人民大学正是抓住这个关键中的关键,自 1994 年开始以"现代汉语述语动词机器词典的研究和建立"为题开展合作,共同承担这个国家自然科学基金项目。一年半来,该项目经过小型的原型实验阶段,进入大规模的工程实施阶段,在原则参数语法指导下,对现代汉语常用的 2300 个动词的 3000 个义项作详细描述,用计算词典方法建造一部信息丰富、结构合理的机器可循词典(Machine Tractable Dictionary),以供从事自然语言理解和语言学研究的学者和科技工作者使用。同时探讨基于大规模语料的计算词典学编辑方法。

二、原则参数语法简介

我们建造"现代汉语述语动词机械机典"所采用的语法理论是当代语法理论中的"原则参数语法"(principles and parameters Approach),也可称为"模组语法"(modular gramar),是"普遍语法"(Universal Granmar,简称 UG)的"核心语法"。为了后面便于说明和讨论,我们在此先对原则参数语法作一简单介绍。

原则参数语法的演进过程前后可追溯三十多年[2][3][4]。乔姆斯基(Chomsky)在 1957 年出版的《句法结构》一书中创立了转换生成语法(transformational generative grammar,简称 TG)新学说,引起了称之为"乔氏革命"的世界语言学界的震动。但 TG 尚未摆脱描写语言学的影响,它忽视语义研究。乔姆斯基 1965 年出版的《句法理论要略》标志着 TG 一个重大发展新阶段,在称为"标准理论"(Standard Theory, ST)的新语法体系中将语义纳入了语法研究范围。标准理论阶段的语义理论是以卡兹—波斯塔假说为基础的,认为跟语义解释相关联的只是深层结构,而由转换而生成的句子的表层结构与语义解释无关。然而随着研究的深入,标准理论的局限性就暴露出来了,这引起六十年代末关于句法和语义关系的大辨论。乔姆斯基在 70 年代初连续发表了《论名物化》、《深层结构、表层结构和语义解释》等几篇重要文章,指出标准理论的不足之处,提出了一个更加精确的语义解释理论—扩充的标准理论(Extended Standard Theory, EST),把 TG 推到一个新的发展阶段。EST 的特点是将语义解释也放到了表层结构,深层结构和表层结构同时为语义解释提供信息。

然而一条语义规则既要对深层结构起作用,又要对表层结构起作用,势必搞得复杂不堪。于是乔姆斯基引进了踪迹理论(trace theory),这样表层结构就可以成为语义的唯一成分。于是TG 进入了"修正的扩展标准理论"(Revised EST, REST)阶段。在这个阶段除了把语义解释全部放到了基层结构上,还在语法中新添了一个层次—逻辑形式。

乔姆斯基 1979 年在意大利比萨的学术讨论会上以演讲形式提出来的"管辖与约束理论"(The theory of Government and Binding, GB)标志着生成语言学的又一次革命,是对ST

(标准理论)的完成。讲稿以"Lectures on Government and Binding"为书名由荷兰福里斯出版社于1981年出版。从此转换生成语法经历着新的转折:研究的重心由原来的规则系统转移到原则和参数系统上。这时产生了普遍语法的概念。实际上乔姆斯基在七十年代中期以后发表的几部涉及语言哲学的著作中都再三强调要对普遍语法进行研究,既使他那些解决具体语法问题的论文也是以研究普遍语法为最终目的[5][3]。乔姆斯基假设:人的语法知识有两部分。一部分是全人类共有的普遍语法知识,这是人类通过生物进化和遗传先天获得的,是不需要学的。另一部分是各民族的语言特有的个别语言知识,这是人们在出生以后通过学习所掌握的。以普遍语法为基础,以经验为外因条件,人才能获得完整的语法知识。

普遍语法"当然不是一部语法,而是一系列条件,限制人类可能有的语言的各种语法的范围"[6]。对普遍语法,乔姆斯基希望[7]:"研究出一套结构高度严谨的普遍性语法理论,它以一系列基本原则为基础,这些原则明确划定语法的可能范围,并严格地限制其形式,但也应该含有一些参数,这些参数只能由经验决定。"管辖与约束理论就是朝这个方向所作的努力。管辖与约束理论的核心就是一系列互相联系、互相制约互动的基本原则(principles)。这些原则具有普遍性,适用于各种语言,同时又具有灵活性,允许不同的语言在一定范围内有些差异,差异所在之处就是所谓的参数(parameters)。

普遍语法由"核心语法"(care grammar)与"周边"(periphery)而成。核心语法由若干互相独立又互相联系互动的"模组"(modules)而成,主要包括:"规则系统"与"原则系统"两种,研究重心是原则系统。原则系统包括"$\bar{X}$—理论(x bar theory)"、"论旨理论(θ—theory)"、"格理论(case theory)"、"管辖理论(government theory)"、"约束理论(binding theory"、"界限理论(bounding theory)"控制理论(control theory)"等七个子系统[8],另外还有一个"空语类(empty category)"。这些子系统是普遍性原则,而且都含有若干数值未定的参数,委由个别语言来选定,故名曰"原则参数语法"。原则参数语法在演进过程中前后有"修正的扩展的标准理论"、"管辖与约束理论"等几种不同名称(后来还有人提出"屏障与约束理论(Barrier—Binding Theory)名称),但乔姆斯基并不赞成随研究重点转移而更改语法理论的名称,并认为"原则参数语法"是最能概括这个语法理论特性的名称[3]。

$\bar{X}$—理论可视为对词组结构本身的合法度条件,把"VP→V comp"、"NP→N comp"、"AP→A comp"等短语结构重写规则合并归成"XP→X comp"形式,其中X为变项,可用V、N、A、P中任一项代入,并且用标准写法"$\bar{X}$",$\bar{X}$表示比X高一层次的语类,比$\bar{X}$更高一层次的语类可用$\bar{\bar{X}}$表示,由此可画出其结构树形图。若把最低层写作X,最高层管作XP,中间需要几层就加几条横线,则XP之下整个树形图是XP所属的最大投射(maximal projection)。

论旨理论要求述语的论元与其"论旨角色"(θ—role)必须是一对一的对应关系;格理论要求具有语音形态(即除了空语类之外)的实号名词组都必须具有"格位"(case);管辖理论是讨论主管成分(中心语)与受管成分(补语)的结构关系的;约束理论是研究语言解释中照应关系的理论;界限理论研究对转换范围的限制;控制理论用来解释句子中的空语类。由于篇幅原因,都不在此详述。

三、现代汉语述语动词机器词典的研究

我们的现代汉语述语动词机器词典的研究和建立有两个基础。第一，中国人民大学语言文字研究最近几年用人工方法对2千多个动词的3000多个义项进行了研究和描写，并据此编著成《动词大词典》[9]。第二，清华大学计算机系在计算语言学的多个领域都有研究成果和经验，并有几千万字的汉语语料。

现代汉语述语动词机器词典(以下简称动词机器词典)的设计思想有以下几个：

1. 以原则参数语法作为理论指导，以论旨网格(theta－gird)方式对每个动词的组合关系从“论元属性(argument)”、“论旨属性(thematic property)”、“句法范畴(category)”(即论旨角色的语类)、“论旨角色的句法功能(syntactic function)”作详尽描写；

2. 从论旨角色语义约束的角度，建立汉语名词性概念的分类体系(thesaurus)以确定名词性概念的聚合关系和上下位关系；

3. 把主要以语言工作者的语感为依据的传统词典学编辑方法同主要以从机贮语料库中获取的大量例证为依据的计算词典学编辑方法结合起来，以使动词机器词典的研究和建立真正立足在丰富和客观的语言事实基础上；

4. 除了构造一部现代汉语述语动词机器词典外，还建立一个功能各全的软件支撑环境。除了对词典作批量录入、单条插入、修改、删除、查询、模糊查询、统计、浏览之外，还考虑探索机器自动编辑出版人用词典的功能。

基于以上的设计思想，清华大学和中国人民大学组成的研究小组在黄昌宁教授、林杏光教授的带领下在陈群秀的组织下对动词机器词典进行研究和实施建立。

国内外对述语动词的论旨关系的研究虽已开展了多年，但像中国人民大学那样对3000个义项作出工程性的具体描写，在国内是首创，在国外也属罕见。但是由于中国人民大学的研究主要凭借语言学工作者的内省，尚需要利用大规模语料来进一步验证和完善。另外，《动词大词典》中只给出每个述语动词的必要论元和论旨角色还不够，现在增加描写每个论旨角色的语义限制、语类和句法功能。因此在此共识之下，研究小组一方面组织人员对二千万语料进行有关动词的句型材料检索，以提供用来人工进行分析归纳，对原有论旨模式进行验证、充实、完善。另一方面，初步建立起汉语名词性概念的分类体系“现代汉语语义分类体系”。在此基础上，设计了《现代汉语述语动词机器词典工作单》，制定了工作单填写规范。工作单包含“词条信息”、“论旨属性”、“其他信息”和“备注”四大部分。词条信息包括“词形”、“拼音”、“动词类型”、“论元数目”、“义项数目”、“义项序号”、“释义”等内容。论旨属性包括“论旨模式”、“论旨名称”、“语类”、“句法功能”、“语义分类”、“语义特征”、“论旨标记”、“论旨实例”等项。而论旨模式又分“基本式1”、“变换式1”、“基本式2”、“变换式2”、“基本式3”、“变换式3”以及“句例”等项目。其他信息包括“否定式”、“时态”、“语义指向动词的后状”、“抽象意义的趋向动词”等。“备注”栏则专门登载“论旨模式的扩展式”。

动词的分类，按照《动词大词典》的分类，将动词分为他动词、自动词、外动词、内动词、系属动词、领属动词六类。

“论元数目”系指该动词的基本论旨模式的必要论元个数，包括域内论元和域外论元。

“义项数目”参照《现代汉语词典》(商务印书馆出版)中的义项数目，如有增加或其他变动，则在该义项数目上打＊号。“义项序号”则是指《现代汉语词典》中该词该义项的序号，若为新增义项，则在该义项序号上打＊号。“释义”也是参照《现代汉语词典》中的释义。如系新增义项或对原《现汉》中释义作过修改，则也标上＊号。

“论旨模式”意指该动词义项下各论旨角色名称与动词的排列顺序(出场顺序)表达式。“基本式1”系指该动词该义项下的最基本的论旨模式,“句例”是该模式的一个或几个句例。“基本式2”、“基式本3”当然是该义项下有别于基本式1的另外两个最重要的论旨模式。

“论旨名称”是指该动词该义项下基本式1、基本式2和基本式3中的必要论旨角色的称呼。开始研究时原准备对《动词大词典》中的22个论旨角色进行扩充、修改和完善(因为原《动词大词典》中22个论旨角色尚觉有些粗泛),后因考虑工作量太大和其他技术原因,而决定不再增加,沿用《动词大词典》中定义的22个格关系(论旨角色)、这22个是“施事”、“当事”、“领事”、“系事”、“受事”、“客事”、“分事”、“与事”、“同事”、“结果”、“基准”、“数量”、“范围”、“工具”、“材料”、“方式”、“依据”、“原因”、“目的”、“时间”、“处所”、“方向”。这22个论旨角色的具体定义参见[10]。

“语类”系指论旨角色的句法范畴,用“{}”将其括起,若同时有几种语类时用“|”表示分隔(表示“或”的意思)。“规范”中规定使用的语类有下列18个:

(1)N:表示名词;
(2)V:表示动词;
(3)A:表示形容词;
(4)R:表示人称代词;
(5)D:表示指示代词;
(6)W:表示疑问代词;
(7)ML:表示数量词;
(8)NP:表示名词词组;
(9)VP:表示动词词组;
(10)AP:表示形容词词组;
(11)S:表示小句;
(12)DE:表示“的”字结构;
(13)N(时间):表示时间名词;
(14)N(处所):表示处所名词;
(15)N(方位):表示方位名词;
(16)N(专名):表示专有名词;
(17)N(普名):表示除N(时间)、N(处所)、N(方位)、N(专名)之外的其他名词;
(18)X:表示任一语类。X=(1)+(2)+(3)+(4)+(5)+(6)+(7)+(8)+(9)+(10)+(11)+(12)。

此外设立N(时间)、N(处所)等N(x),是因为考虑到论旨角色中常常要用到某个N(x)或某几个N(x),分细一些填写准确一些方便一些,而许多情况下又是一切名词,所以有“N”又方便。而设立“X”完全是为了某些动词的语类可为任一语类,但书写|N|V|A|R|D|W|ML|NP|……太麻烦,所以设一个X以概括之。

“句法功能”系指该论旨角色在句子中充当的句法成分。本“规范”采用传统的6种句法成分“主、谓、宾、定、状、补”,实际上在本词典规范中用得上的是主、宾、状、补四个。

“语义分类”意指某论旨角色在语义分类体系中的上位义类。本规范使用的语义分类体系是在陈群秀原拟定的“汉语语义分类体系”基础上修改完善而成的,是在参照国内外义类体系或同义词词典基础上拟定的。书写时用“{}”将其括起,同时有几种语义分类时用“|”表示分隔。若论旨角色中需包含词本身时,用“;”表示词的分隔。亦即可以是语义分类上位与词本身混合表示。此外设立一个Y,表示任一语义分类,意即Y=超类+事+物+时空+部件。

“语义特征”栏的设置是因为有时光填写论旨角色的语义分类还不能限制住,则把语义特征作为辅助性手段。例如“十片状”、“十液体”、“一毒性”等。有了“语义分类”、加某些具体词再加上语义特征,则语义限制就好表示多了。

“论旨标记”系指论旨角色所带的前置或后置的介词或方位词，例如：“在”，“对”，“为”。

“论旨实例”意指论旨角色的实例，用以对论旨角色的“语类”、“语义分类”作形象化的说明或补充说明。

“其他信息”中的各项（否定式、时态、语义指向动词的后状、抽象意义的趋向动词等）则是采用详细列举用打“√”方式填写。其中“语义指向动词的后状”是指“有可能充当动词补充成分的动词”和“有可能充当动词补充成分的形容词”。

“备注”栏填写的扩展式，意指除基本式中必要论旨角色之外的包含非必要论旨角色的扩展式中所含的非必要论旨角色。

以上的“其他信息”及“备注”是一些辅助信息，而“词条信息”中的“论元数目”和“论旨属性”中的“论旨模式”、“论旨名称”、“语类”、“句法功能”、“语义分类”、“语义特征”、“论旨标记”、“论旨实例”等组成“论旨网格”对动词进行详细的主要描述。

四、结语

现代汉语语动词机器词典的联合研究小组在对200个动词作了原型实验之后，对大批动词用2千万语料抽句例，进入了3000词条的大规模的工程实施阶段，并初步建立了动词词典的软件支撑环境。我们最大的愿望是能建立好这部机器词典，提供给从事自然语言理解和处理的科技人员作为他们开发他们系统或进行基础理论研究时的机贮资源，提供给从事汉语语言学研究的科学工作者作为他们研究汉语的有力工具和助手。

参考文献

[1] 陈群秀，黄昌宁，程红，现代汉语述语动词机器词典初探，《计算语言学研究与应用》，北京语言学院出版社，1993年10月。

[2] 陆致极，《计算语言学导论》，上海教育出版社，1990年。

[3] 汤廷池，普遍语法与汉英对比分析，《汉语词法句法续集》，PP. 213－256，台湾学生书局，1989年12月初版。

[4] 汤廷池，国语里“移动α”的逻辑形式规律，《汉语词法句法论集》，PP. 401－448，台湾学生书局，1988年3月。

[5] 徐烈炯，管辖与约束理论，《国外语言学》，No. 2，1984年，北京。

[6] Chomsky, N. (1980), Rules and Representation, Columbia, New York.

[7] Chomsky, N. (1981), Lectures on Government and Binding, Foris, Dordrecht.

[8] Chomsky, N. (1982), Some Concepts and Consequences of the Theory of Government and Binding, linguistic Inquiry Monograph 6, MIT Press, Cambridge.

[9] 林杏光（审定）、鲁川（主编）、王玲玲（副主编），《动词大词典》，中国物资出版社，1994年2月。

[10] 林杏光，进一步深入研究现代汉语格关系，《计算语言学研究与应用》，北京语言学院出版社，1993年10月。