汉语词的概率语法属性描述

俞士汶　段慧明　朱学锋

(北京大学计算语言学研究所 北京　100871)

[摘要]

“现代汉语词的语法属性研究”是中国国家哲学社会科学基金在“九五”期间支持的语言学科重大课题“信息处理用现代汉语词汇研究”中的一个子课题，本文首先介绍这个子课题的主要研究内容，即以《现代汉语语法信息词典》的已有成果为基础，以大规模真实语料的统计数据为依据，用概率值重新描述词的语法属性。然后介绍这个子课题已经取得的阶段性成果，并探讨进一步发展的方向。

[关键词] 现代汉语，词汇，词类，词的语法属性，概率语法属性描述
[中图分类号]　　[文献标识码]　　[文章编号]

Probable Grammatical Attribute Description of Chinese words

Yu Shiwen,　Duan Huiming,　Zhu Xuefeng

Abstract: Based on the Grammatical Knowledge-base of Contemporary Chinese developed by the Institute of Computational Linguistics, Peking University (ICL/PKU) and the Large Scale Tagged Corpus jointly developed by ICL/PKU and Fujitsu, the Grammatical Attribute of the Chinese words will be re-described with statistical data acquired from the corpus. The paper introduces this new project, some results and further researches.

Keywords: Contemporary Chinese, Lexicon, Parts Of Speech, Grammatical Attribute of Words, Probable Grammatical Attribute Description

1. 词的概率语法属性的提出

笔者认识到汉语自动分析的一些特殊困难[1,2,7],并认为克服这些困难的必要手段之一就是建立综合型语言知识库[3,4]。词的语法属性知识是这个知识库的重要组成部分。北大计算语言学研究所与中文系长期合作，已研制了一部用于信息处理的《现代汉语语法信息词典》[5,9]。这部词典已收词语 7.3 万余条。在建立了基于语法功能分布的分类体系后，完成了7.3 万词语的归类，并进一步对各个词类中的每一个词语详细描述它们的多项语法属性信息。这部词典已在国内外几十个单位的语言信息处理研究中发挥作用。一部详细介绍这部词典的专著于 1998 年出版[5]，清华大学出版社拟在 2001 年内出版第二版。

这部词典中的语法属性信息的值多数为“可否型”或“是非型”，在此基础上建立语法规则，仍然难免“说一不二”或“非此即彼”， 缺乏柔性。“现代汉语词的语法属性研究”子课题就是以《现代汉语语法信息词典》的已有成果为基础，以大规模真实语料的统计数据为依据，用概率值重新描述词的语法属性，这是一项全新的研究工作。

2. 词的语法属性值的类型

在语法信息词典中,描述词语语法属性的字符型属性值划分为“二选一型”、“多选一型”、“复合型”和“释义型”4 种子类型。除“释义型”外，其他 3 种属性值都可以按不同方式用概率值替换。

2.1 二选一型

这种属性值只有两种可能的选择，最便于替换。比如：动词有一个属性描述其能否受程度副词修饰。像“想、重视”这些动词能受“很”修饰，可设值为“Yes”； 像“等于、调查”等动词不能受“很”修饰，设值为“No”。这种类型又可叫做“可否型”。 在语法信息词典中，这种类型最多，相当于数据库中的逻辑型，逻辑型字段取值要么是“真”，要么是“假”。“可否型”属性值的实际含义有所不同，“No”是刚性的，“Yes”是弹性的。如果某动词的“很”这个属性的值为“No”，那么该动词一定不能受“很”修饰；如果其值为“Yes”，只是说明该动词可以受“很”修饰。但并不指明该动词在实际使用时受“很”修饰的这种可能性有多大。

如果将某种属性看作随机变量，用随机变量的概率值刻画该属性，则既可以客观反映语言的模糊性，又可以避免个人语感等因素的干扰。仍以动词的“很”属性为例。设某动词在语料库中共出现 m 次，其中实际受“很”修饰的有 n 次，则定义该动词受“很”修饰的概率值为

$$p=n/m,$$

以“p”作为该动词的“很”属性的值，就完成了对动词是否受“很”修饰这个属性的描述方式的改造。显然，这样的描述更为科学，更为客观。

适应某些应用的实际需要，反过来又可以在概率属性值的基础上重新建立“可否型”的描述方式。设一阈值 δ。若 $p \geq \delta$，则定为“Yes”；若 $p < \delta$，则定为“No”。

在“可否型”的属性值中还可以更精细地区分出“是非型”。“是非型”是真正的逻辑型，如在动词库中有这样的一些字段：“系词”、“助动词”、“趋向动词”、“形式动词”等。这些字段指明每个动词是不是系词、助动词、趋向动词、形式动词。从逻辑角度考虑，“是非型”字段无需用概率值改造。

2.2 多选一型

动词库中的“体谓准”属性字段是多选一型的，其值可定义为以下 8 种，即“内”、“体”、“谓”、“准”、“体谓”、“体准”、“谓准”、“体谓准”，分别代表该动词是不及物动词、只可带体词性宾语的动词、只可带谓词性宾语的动词、只可带准谓词性宾语的动词、可带体词性宾语和谓词性宾语的动词、可带体词性宾语和准谓词性宾语的动词、可带谓词性宾语和准谓词性宾语的动词、可带体词性宾语和谓词性宾语以及准谓词性宾语的动词。为了用概率值描述这些属性，将这一个字段拆分为 3 个字段：“体宾”、“谓宾”、“准谓宾”，并规定这 3 个属性值的类型都是数值型的。设某动词在语料库中以 v 的形式出现 m 次，实际带体词性宾语的有 x 次，实际带谓词性宾语的有 y 次，实际带准谓词性宾语的有 z 次，则定义这 3 个属性的概率值分别为：$p1=x/m$，$p2=y/m$，$p3=z/m$. 若 $p1+p2+p3=0$，则该动词是不及物动词的可能性很大。

2.3 复合型

本来在关系数据库理论中规定所有字段的值必须是“原子”,即字段的值是不可再分割的，

不过，实用的关系数据库管理系统都可实现这样的附加操作，即从字符型字段的值中取出其中的子字符串。这里将可分解成若干子字符串的类型叫做复合型。在名词库中有“ 个体量词、度量词、容器量词 ”等字段。如对于“ 白菜 ”，其个体量词可填“ 棵，个 ”，其度量词可填“ 斤，克，千克，公斤 ”，其容器量词可填 “ 筐 ”，其种类量词可填 “ 种 ”，其成形量词可填 “ 堆 ”，其不定量词可填“ 些，点 ”。同样，量词库的“ 后名 ”字段的值也是复合型的。如对于“ 杯 ”，“ 后名 ”字段可填“ 水，茶，酒，咖啡 ”；对于“ 本 ”，“ 后名 ”可填“ 书，杂志，小说 ”。

对于这类复合型字段，改用概率值描述的方法有所不同。每个具体的量词不便作为名词库的字段名称，因为汉语中常用量词有数百个。而将所有的名词作为量词库的字段更不现实。因此，将不改变这类复合型字段的值的类型。

从实际语料中可以统计“ 白菜 ”受“ 数量短语 ”修饰的次数和不同量词的使用次数。设“ 白菜 ”在文本中出现 m 次，其中受含“ 棵 ”的数量短语修饰的有 n1 次，含“ 个 ”的有 n2 次，含“ 斤 ”的有 n3 次，含“ 公斤 ”的有 n4 次，含“ 筐 ”的有 n5 次，含“ 堆 ”的有 n6 次，含“ 种 ”的有 n7 次，含“ 点 ”的有 n8 次。令

$$pi=ni/m,\quad (i=1,2,3,4,5,6,7,8)$$

pi 则为上述 8 个量词与“ 白菜 ”搭配的概率。可以考虑将“ 白菜 ”的有关量词的各字段作如下改造。

字段名称	字段的值
个体量词	棵:p1,个:p2
度量词	斤:p3,公斤:p4
容器量词	筐:p5
成形量词	堆:p6
种类量词	种:p7
不定量词	点:p8

为了能整体地考察名词受数量短语修饰的情况，可另增加用概率值描述的字段。例如，可增加一个“ 数量短语修饰概率 ”字段，对于“ 白菜 ”，此字段可填以

$$p=n/m,$$

这里，

$$n=n1+n2+n3+n4+n5+n6+n7+n8.$$

即说明白菜受数量短语修饰的概率是 p．若某个名词的 p=0,则说明该名词受数量短语修饰的概率是 0，那么，认为该名词是无量名词的可信度则更高。

3． 基础资源的准备

3.1 早期规划的数据资源

从上述概率语法属性模型可以看出，本子课题的研究工作依赖语料库。语料库当然越大越好。不过资源总是有限的。笔者以为当前语料库有 2,000 万字的规模基本可以满足需要。语料的题材和体裁要相对平衡。考虑到《现代汉语语法信息词典》的通用性，拟建的“ 词的概率语法属性库 ”也是面向通用目标的。因此，国家语委主持开发的“ 现代汉语平衡语料库 ”覆盖当代汉语的部分有 2,000 多万字，是适合需要的。

本子课题的另一个必要资源是“ 现代汉语基本词表 ”。该词表应以大量语料的统计为基础，包含 20,000 个左右使用频度最高的词语。本子课题曾打算以这样的词表为对象建立“ 词的概率语法属性库 ”。

本子课题需要对语料库进行最基本的词语切分和词性标注的加工。将 2,000 万字的语料实现正确切分并标注词性是件大工程。 笔者曾认为在本课题的限期内，还不会有这样的语料

库。因此，考虑了针对性更强的加工方案：在语料库中对基本词表建立文中关键词索引 KWIC（ Key Word In Context），并在关键词的两侧截取一定长度的上下文作为研究基本词的语法属性的环境，左侧和右侧取的长度可以不等，随所考察的语法属性而变动。

进一步的针对性加工则取决于词的属性研究的目标。例如，要研究动词带宾语的情况，则要标注出直接跟在该动词后的宾语，并标注该宾语的性质（体词或谓词或体词性短语或谓词性短语或小句等）。

在课题进展期间，北大计算语言学研究所拥有的标注语料库已达到足够大的规模，就没有将上述针对性的加工任务付诸实施。但这里提供的设计思想在今后的研究中仍是有借鉴意义的。

3.2 实际利用的语料库资源

北大计算语言学研究所与 FUJITSU 合作，对《人民日报》1998 年全年的语料（2600 万字）进行加工。目前的加工项目包括词语切分和词性标注，并标出专有名词（包括短语型专有名称）[7,8]。经过如此加工的语料库可以简称为“ 标注语料库 ”。以下示例是从标注语料库中摘录的。

由/p [共青团/n 中央/n]nt 、/w [全国/n 绿化/vn 委员会/n]nt 、/w 林业部/nt 、/w 铁道部/nt 、/w [全国/n 青年/n 联合会/n]nt 共同/d 发起/v 的/u 迎/v 香港/ns 回归/v 京九/j 植绿护绿/l 活动/vn 今天/t 正式/ad 启动/v 。/w 广东/ns 的/u 深圳/ns 、/w 惠州/ns 、/w 河源/ns 等/u 地/n 同时/d 举行/v 了/u 隆重/a 热烈/a 的/u 启动/vn 仪式/n 。/w

加工后的语料，切分单位（词语）之间用“ 空格 ”隔开了。n,v,a,w 分别表示它左边的切分单位是名词、动词、形容词、标点符号等，ns 是地名，nt 是团体机构的名称。方括号中的内容代表一个短语型专有名称。加工依据是《现代汉语语料库加工——词语切分与词性标注规范》[8]，其中词性标记与语法信息词典的 26 个词类代码一致[5]，另增加了以下 3 类标记：①专有名词的分类标记，即人名 nr，地名 ns，团体机关单位名称 nt，其他专有名词 nz；②语素 g 按其子类标注，已有名语素 Ng，动语素 Vg，形容语素 Ag，时间语素 Tg，副语素 Dg 等；③动词和形容词的某些功能标记，即名动词 vn（在句法结构中起名词作用的动词），名形词 an（起名词作用的形容词），副动词 vd（作状语的动词），副形词 ad（作状语的形容词）。合计约 40 个左右。

《人民日报》的纯文本文件的质量高，几乎没有错字。1998 年一年的语料虽然不能说已全面覆盖了当代汉语，但至少是一个相当大的有代表性的子集。到 2000 年 6 月，半年语料的加工任务已经完成。1300 万字的标注语料库为本子课题提供了相当丰富的合适的数据资源。

对《现代汉语语法信息词典》中的语法属性字段进行筛选，制订了各类词语的可直接统计的语法属性表，并在经过严格校对的 1300 万字的标注语料库上进行统计，取得了有价值的成果。

需要说明一点，本子课题在 2001 年 3 月结题前能利用的语料库只有 1300 万字，预料会有数据稀疏问题。因此目前对大多数属性只给出统计量，不计算概率值。一年后便有 2600 万字的资源，那时，多数语法属性的概率值的计算将是水到渠成的事。

4．成果概要

4.1 带词性的词频统计

这是汉语学界首次取得的成果。我国做过词频统计工作[10,11]，但没有在千万字量级的

语料库上做过带词性的词频统计。现在，不仅得到了 6 个月所有语料上的统计数字，而且可以得到每个月的统计数字。因此可以了解词的使用频度随时间变化的情况。只有那些不仅一年期间的总频度高，而且按月分布均匀的词才真正是最常用的词。结果显示，助词“ 的 ”、介词“ 在 ”、连词“ 和 ”、助词“ 了 ”、动词“ 是 ”、数词“ 一 ”、副词“ 不 ”、动词“ 有 ”、介词“ 对 ”、专名“ 中国 ”是最常用的前 10 个词。“ 中国 ”居第 10 位反映了《人民日报》的特点。

还可以得到同形兼类词的分布情况。例如，“ 把 ”单独作为一个词共出现了 10,221 次，其中介词 9,801 次，量词 284 次，动词 112 次，数词 18 次，名语素 6 次。为了检验“ 把 ”作为数词的用法，从语料库中检索例句，得到“ 这条河有百把里长 ”这样的例句，这里的“ 把 ”确实是助数词。

还能导出词类的频度。在语法信息词典划分的各词类中，名词 n 的频度最高，其次为动词 v。出乎意料的是标点符号 w 居第 3 位。

4.2 词性转移矩阵

在计算机上实现语言统计模型至少需要 2 组数据。其一是上述的带词性的词频。其二是在已知前一个词性的条件下，某个词性出现的概率[6]。词性转移矩阵正是这组数据。早在 20 世纪 90 年代，已有过这方面的研究[12]，但标注语料库的规模很小。截止到目前为止，其他的中文标注语料库的规模也都只有几百万字。现在，笔者在 1300 万字标注语料库上得到了“ 词性转移矩阵 ”这样有意义的数据。从这个矩阵可以查到，前一个是副词，后面出现了 39,464 次形容词；反过来，前一个是形容词，后面只出现 1,772 次副词。需要注意，形容词后面出现副词，并不等于说这个“ 形容词+副词 ”序列是在同一个短语结构层次中。

4.3 《现代汉语语法信息词典》中可统计的属性清单

如对于时间词、处所词、方位词，可以从标注语料库中直接统计它们在“ 到 ”、“ 在 ”、“ 自从 ”的后面出现的次数。还可以进一步联合使用语料库和语法信息词典。从目前的标注语料库，只能笼统地得到名词前面出现量词的次数，结合语法信息词典中的量词子类信息，就可以更细致地得到名词前面分别出现各类量词的次数。全面考察了语法信息词典的所有属性字段，列出了各类词语的可统计的属性清单。

4.4 词语属性的统计数据

对 1300 万字标注语料库中的可入词典的所有词语（像“ 1998 年 ”、“ 1449 ”等时间词、数词通常不收入词典），按**可统计的属性清单，**逐一进行了统计。例如，按动词的“ 受‘ 很 ’修饰 ”这个属性进行统计，对“ 想 ”得到的结果为“ 12 ”，即“ 想 ”有 12 处受“ 很 ”直接修饰；对“ 吃 ”、“ 到 ”、“ 发展 ”、“ 进行 ”、“ 走 ”等皆为“ 0 ”，这些结果验证了《现代汉语语法信息词典》中的属性值的可靠性。

不过，也要注意，目前是在二元语法模型的基础上进行统计，未涉及短语结构。通常认为“ 有 ”不能受“ 很 ”修饰，但对“ 有 ”统计受“ 很 ”修饰的情况，得到的结果却是 141 次。对这 141 次的合理解释是“ 有 ”虽然不能直接受“ 很 ”修饰，但带了某些宾语后又可以受“ 很 ”修饰的，像“ 很有意思 ”、“ 很有勇气 ”、“ 很有水平 ”都是正确的。由于只基于二元语法模型，“ 有 ”前面出现了“ 很 ”，也只好统计进来。“ 有 ”是第 8 个高频词，频次为 30,298，只有 141 个前面有“ 很 ”，概率值是小的。又例如，对于“ 三双尼龙袜子 ”，若加工为“ 三/m 双/q 尼龙/n 袜子/n ”，统计时认为“ 尼龙 ”前面有量词“ 双 ”，而不能正确地判定这个“ 双 ”实际上是修饰名词短语“ 尼龙袜子 ”的中心语“ 袜子 ”的。

5. 进一步的研究

本子课题所取得的成果可以在语言信息处理的实际系统中得到应用，也可以为语言教学提供宝贵的资源。

显然，需要进一步研究的课题还很多。数据稀疏问题明显存在。期望有了 2600 万字的资源后，能得到改善。不过，语料平衡问题还会存在。短语标注是提高词的语法属性统计数据正确性的关键。其他像义项标注、注音都是极有价值的工作。在大规模语料库加工的漫长征途上，北大现在做的工作还只是刚刚迈出了第一步。

6．致谢

感谢国家语委科研规划领导小组办公室与许嘉璐教授将国家社科基金语言学科“ 九五 ”重大课题“ 信息处理用现代汉语词汇研究 ”之子课题“ 词的语法属性描述研究（97@yy001-6）”分配给北京大学计算语言学研究所。在许嘉璐教授的部署与指导下，经过两年多的努力，子课题组取得了若干成果。更有意义的是，在这个重大课题的带动下，北大计算语言学研究所建立了多个新的研究项目，拓广了研究视野。

除笔者外，本子课题组的成员还有王惠、张化瑞。

陆俭明、胡明扬、冯志伟、詹卫东等先生对本课题的开展提供过有价值的建议。本子课题组全体成员表示衷心的谢意。

参考文献

[1] 俞士汶，关于受限的规则汉语的设想，见王均主编《语文现代化论丛》，山东教育出版社，1995 年 10 月，193-205

[2] 俞士汶，朱学锋，受限汉语研究的必要性，见王均主编《语文现代化论丛第三辑》，语文出版社，1997 年 10 月，150-160

[3] 朱学锋，俞士汶，自然语言处理与语言知识库，见罗振声，袁毓林主编，《计算机时代的汉语汉字研究》，清华大学出版社，1996 年，107-118

[4] 俞士汶，关于语言信息处理技术的展望，《计算机世界》，1997 年第 1 期，第 127 版

[5] 俞士汶，朱学锋，王惠，张芸芸，现代汉语语法信息词典详解，北京：清华大学出版社，1998 年 4 月，第 1 版

[6] 周强，基于语料库和面向统计学的自然语言处理技术介绍，《计算机科学》， 1995，22（4），36-40

[7]段慧明、松井久仁於、徐国伟、胡国昕、俞士汶，大规模汉语标注语料库的制作与使用，《语言文字应用》，2000 年第 2 期，72-77

[8] 俞士汶、朱学锋、段慧明，大规模现代汉语标注语料库的加工规范，《中文信息学报》，第 14 卷，第 6 期，2000 年，58-64

[9] 俞士汶、朱学锋、王惠，《现代汉语语法信息词典》的新进展， 《中文信息学报》，第 15 卷，第 1 期，2001 年，P58-65

[10]北京语言学院语言教学研究所，《现代汉语频率词典》，北京 :北京语言学院出版社，1986 年 6 月，第一版

[11]刘源等著，《信息处理用现代汉语分词规范及自动分词方法》，北京：清华大学出版社，1994 年 6 月，第一版

[12]白栓虎等，汉语语料库词性标注方法研究，见 :陈肇雄主编《机器翻译研究进展》，北京：电子工业出版社，1992 年 8 月 1 版

[作者简介]俞士汶，男，教授；段慧明，女，高级工程师；朱学锋，女，副教授；3 人皆从事计算语言学研究，属北京大学计算语言学研究所。通信地址：中国 100871 北京市，北京大学计算机系。电话：86-10-6275189，传真：86-10-62756591，Email: yusw@pku.edu.cn，

duenhm@pku.edu.cn, 主页：www.icl.pku.edu.cn 或 icl.pku.edu.cn