汉语义频词库的开发研究

[本栏导读] 在计算机的语言处理中实现词义自动标注,是当前中文信息处理界关注的一个热点问题。国家社科基金“基于国家语委‘通用语料库’的汉语义频词库的开发”(项目编号:04BYY009,负责人:苏新春)课题组立足于大规模语料库,从语义、语法层面对 3771 条多义词义项的区别性形式特征进行了系统描绘,构建了一个包括词语库(义项库)、语料库、语义库、规则库、规则验证库、语法—语义分析框架等多个子系统的“多义词词义搭配知识库”,通过语料处理平台、搭配规则发现平台、词义标注系统来实现词义的自动识别。本期刊登的 5 篇论文,分别对词语库的收词与规则库的建构原则、机用词典义项的形式特征、多义词义项语义关系及对词义消歧的影响、动词与形容词的搭配特征等问题作了深入探讨,多角度介绍了该知识库的建构理论、方法与实践。

词语库的收词与规则库的建立

苏新春　杜晶晶

(厦门大学　福建　厦门　361005)

[摘要] 词语库与规则库是在“多义词词义搭配知识库”中起基础与核心作用的两个子库。词语库有两个来源,一是词典词,二是真实语料词,两类词语有着书面语与口语词、正体词与异体词、语言词与言语词、通用词与领域词、稳定词与具体词等方面的差异。词语库特点会在很大程度上影响到词义标注的效果与正确率。纳入首批考察的词语为双音节多义词 3771 条,共有义项 7861 个。规则库统摄语义库、义项库、语料库,这些知识库通过规则库的组织而发挥作用。规则库是实现词义标注工程目标的直接依据,对于任何一个多义词,规则定义的多寡有无、质量好坏都会直接影响标注结果。规则库集中体现 SCT 整个系统的意义与价值,是语言知识与工程实施的结晶体。

[关键词] 多义词; 词义搭配知识库; 词语库; 规则库

[中图分类号] H08　**[文献标识码]** A　**[文章编号]** 1003-5397(2014)01-0010-10

[收稿日期] 2013－12－02

[作者简介] 苏新春,厦门大学中文系教授,博导,主要研究汉语词汇、辞书语言、教材语言; 杜晶晶,厦门大学中文系博士生,嘉庚学院人文与传播学院讲师,主要研究计量词汇学、国际汉语教学。

The Word-Collection of Word Base and The Establishment of Rule Base

SU Xinchun, *DU Jingjing*

Abstract: Word base and rule base are the two central and fundamental subsets of The Polysemy Sense Collocation Knowledge Base. Word base is comprised of words from dictionaries and words from corpora. They differentiate in written words and spoken words, standard words and variants, language words and speech words, general words and domain words, stable words and concrete words, etc. The characteristics of word base will to a large extent affect the result and accuracy of word sense tagging. The first study includes 3771 polysemou disyllables with 7861 word senses. The semantic base, sense base and corpus are subject to the organization of rule base. Rule base is a direct basis for word sense tagging. For any of the polysemous words, tagging depends on the quality and quantity of rule definition. It also epitomizes the significance and value of the whole SCT system, and is the combination of linguistic knowledge and word sense tagging.

Keywords: polysemous words; word sense collocation knowledge base; word base; rule base

"多义词词义搭配知识库"(SCT)是为了实现词义自动标注供计算机运行的一个语言知识系统。在这个系统中有两个子库起着基础与核心的作用,那就是词库与规则库。词库是词义自动标注的对象,里面最有价值的是多义词。规则库是词库在多种资源库的支持作用下,把不同义项的形式特征以规则的方式予以标示,并登记入库的标示集。多种资源库包括语义分类库、语法分析框架、测试用的标注语料库、校正用的人工预估的义频库等,它们都是为了帮助识别多义词的不同义项而在不同层面、不同角度起着辅助作用。本文主要对词库与规则库这两个基本对象库的功能与设计原则进行论述。

一　词库的收词特点

词库收录的是用于词义自动标注的词。这看上去似乎不是问题,收多少词,收什么样的词,不就是把需要标注的词特别是多义词都收进去吗?可细心琢磨,问题并不那么简单,无论是词典中的词还是源于真实语料的词,都存在许多需要认真对待的问题。

(一)来自词典的词

反映现代汉语词汇面貌的辞书,《现代汉语词典》(以下简称《现汉》)无疑最有代表性。首先,它是一部语文词典,收录的主要是人们日常使用的生活词语。其次,它是一部中型词典,收录了 6 万余条词语,现代汉语的基本词、常用词都在其中。最重要的是,它的功能就是为了反映现代汉语词汇的构成与概貌。"《现汉》的编纂者多是造诣精深、学有专攻的行家里手,他们从上百万张资料卡片中反复斟酌,层层筛选,最后确定收录的五万多条词语,无疑是对现代汉语词汇的一次全面整理和规范。"[1] 但在使用中发现,《现汉》也存在着不甚吻合、甚至抵牾的地方。这里姑且不说义项的分割是否合理,相互之间是否清晰,就是词语与义项的有无上也存在着不少问题。

1. 真实语料有而词典未收的词

词典的收词永远落后于语言实际生活，这是永恒的现象。如“制片”“迪厅”“黑屏”“恶搞”，在生活中已较常见了，可一般的语文词典仍未收录。这里说的还远不是这种情况，而是指传统词典有意回避而不予收录的“固定结构词”。“所谓固定语就是它们不像一般的词语那样有独立运用功能较强和意义较完整的特点，只在言语使用中常常紧密地结合在一起。”[2] 这样的词语一般也不为“规范词典”所关注。如“不能”“亏心事”在实际语料中并不少见，可词典却未收录：

(1) 想到/v 自己/r 竟然/d 不能/v 得到/v 这/r 姑娘/n 的/u 感情/n ，/w 想到/v 自己/r 的/u 孤单/an ，/w 心里/s 又/d 委屈/a ，/w 又/d 凄凉/a ，/w 也/d 不禁/d 流下/v 眼泪/n 。/w

(2) 我/r 没/d 做/v 那/r 亏心事/n ，/w 我/r 还/d 觉得/v 我/r 挺/d 有/v 道理/n 哩/y！”

要弥补这个不足就要给词库尽量添置多的词，否则词库的规模过小，词库的词与实际语料中的词语数相差过大，词义标注的面必然狭窄。词典之所以会有这种情况，是因为传统词典对词语的收录有着诸多限制，反映实际的“词汇面貌”只是其功能之一，而注重查考性，排斥见字明义者，乃是其沿续长久的一个收词传统。

2. 真实语料有而词典失收的义项

真实语料中有的词义，而在多义词的词库中并没有。这是来源词典义项的另一不足。如“安静”。《现汉》收的两个义项是“①没有声音；没有吵闹和喧哗。②安稳平静。”《现代汉语规范词典》收的两个义项是“①没有声音；不嘈杂。②安稳平静。”《当代汉语词典》收的两个义项是“①没有声音；没有吵闹和喧哗。②安稳；平静。”

在真实语料中却有着大量这样的用例：

(3) 在这首歌中，流行音乐、戏曲音乐和民歌的手法和素材交融，音乐紧紧扣住在演唱者最富有色彩的中音区，使音乐在平缓、安静和从容中流淌。

(4) 那些德高望重的文化人，一般都很安静、很严肃，保持着一种过来人的平和，无声无息地往那里一坐，就让你肃然起敬。

(5)“围棋讲究潜心钻研，一躁一动再加上点满足，那份安静的心境自然会变了调。”

显然，这里的“安静”表示的是“沉静稳重”义，以上三部词典的两个义项并不能准确地概括。显然应该在词库中增加这一义项。

又如“打印”有两个义项“①打字油印。～盖章。②盖图章。”(《现汉》第 3 版) 这两个义项不能涵盖下面的例子：

(6) 就因为她有一本不具名的打印的诗集。

《现汉》第 5 版为“打印”新设立了一个义项“把计算机中的文字、图像等印到纸张、胶片等上面。”这是传统词典“与时俱进”的结果。因此，在利用词典的多义词材料时，就一定要保持着这份警惕，时刻注意把缺漏、新生的义项补进去。

3. 真实语料库有而词典故意不收的义项

有一类义项是传统词典有意不收的，即“见字明义”的日常生活义。如“不平”收了下面 4 个义项“①不公平。②不公平的事。③因不公平的事而愤怒或不满。④由不公平的事引起的愤怒和不满。”没有收“不平整、不平滑”义，而恰恰正是这个义项大量存在于实际语料中。如：

(7) 原来不平的脸部皮肤，又变得平整光滑，就像少女时代一般。

显然,这样的常用义是应该收进面向真实语料的词库的。

(二) 来自真实语料的词

1. "通用语料库·核心库"

我们使用的真实语料是国家语委的"通用语料库·核心库"。研制者对研制过程与原则有过完整说明[3]。从说明中可知这是一个具有综合性、通用性、平衡性、断代性的大型语料库。把它作为现代汉语共时状态的反映,作为提取现代汉语通用词语的来源是有代表性的。我们对"核心库"的语料的规模与类型做了调查:

1) 规模与容量: 共 9487 个语料样本。样本最大为 11,124 字,最小为 153 字,平均每个样本 2151. 5 字。

2) 时代分布

20 世纪 80 年代的语料占 1/2 略多,90 年代(主要是 1990 ~ 1993 年,另外 1996 年 1 篇,1998 年 6 篇)约占 1/4。具体见表 1。

表 1　"核心库"语料时代分布

时间	样本数	百分比
20 年代	15	0. 16
30 年代	488	5. 14
40 年代	249	2. 62
50 年代	596	6. 28
60 年代	452	4. 76
70 年代	498	5. 25
80 年代	4955	52. 23
90 年代	2221	23. 41
时代不明	13	0. 14
总数	9487	100

3) 语料类型分布见表 2。

表 2　"核心库"语料类型分布

语料类型	数量	百分比
人文与社会科学	7828	82. 51
自然科学	1459	15. 38
综合类	200	2. 11
总数	9487	100

2. 从"核心库"提取词表存在的问题

即便是具有如此普遍性、通用性的语料,从中提取的词表仍有需要进一步完善的地方。

1) 分词的差错

"词"的切分准确与否,会直接影响到词汇的数量与分布。请看下面的例子:

(8) 用/v 压/v 电/n 材料/n 做小/v 平面/n 镜/n 阵/n 来/vd 代替/v 一块/d 反射镜/n ,/w 每/r 块/q 小/a 平面/n 镜/n 可以/vu 自动/a 调节/v ,/w 或者/c 把/p 主/n 镜/n 设计/v 得/u 可以/vu 快速/a 改变/v 其/r 局部/n 的/u 形状/n ,/w 以/p 在/p 最后/n 的/u 焦/a 平面/n 上/nd 获得/v 消除/v 大气/n 湍/v 动/v 和/c 光学/n 像/n 差/a 影响/v 的/u 天体/n 像/n 。/w

共分出 53 个分词单位,里面就有若干值得商榷的地方:

A. 可分可合,分开后专业词语消失了。如"压电""镜阵""焦平面""像差""天体像"都被分开了。它们合起来是专业词,反映了专业词语的面貌,分开则只剩下"通用"的语文词了。

B. 不该分的语文词分了。如"湍动"被分开,这个词就消失了。

C. 不该合的合了。"做小"是一个因结构误切而新产生出来的。它在原文中可以轻松地被发现是误切,可一旦列入词表却不会被发现,因为语言生活中是有"做小"这样一个词的,意思是"低声下气""做小老婆"。在核心库的词表中"做小"有两例,都是误分的结果。

这三种情况都会造成词语汇总表的正确性。A 反映不出词汇特点,B 减少了词语的种数,C 造成错误的词频数据。这些提取出来的词语当然都会直接影响到词义标注的效果。

这样的误切误分现象不少。如:

行业词的消失: 如“白—磷”“滤—纸”“X—光—室”“卫—线”“信—徒”“互济—会”“行为—科学”。

专名的消失: 如“一顶‘文艺—黑线—专政—论’的帽子”“总结出‘傻—论’者的观点”“还给它起了一个吓人的名字‘狼—桃’”。

言语词的消失: 如“千万不要认为我这县长能‘一—掌—遮—天’”;“现在不是提倡‘访—富—问—甜’吗”;“这一席话,‘言—简—情—深’”;“先定一个框框,拿框子去套,接着就是抓辫子,挖根子,戴帽子,打棍子,那就不好了嘛。一来就是‘五—子—登科’”;“目—不—旁—视”。

新词语的消失: 如“‘剧—画’就是我国的连环画”。

复合词的消失: 如“其燃料多用‘枣—木炭’”。

那么对分词错误应该做出怎样的一个估计呢? 分词错误当然会直接影响到词语统计与词义标注的正确性。对差错率的估计,其实就是对容错度的估计。下面是“国人”一词的统计。核心库中的词频共有 71 例,错分的有 16 例。如:

(9) 过去/nt 几/m 年中/nt 两/m 国人/n 民间/n 频繁/a 的/u 交往/v 。/w

(10) 大致/d 有/v 二/m 种/v 情况/n :/w 一/m 是/vl 法律/n 对/a 我国/n 公民/n 的/u 适用/v 范围/n ;/w 二/m 是/vl 法律/n 对外/d 国人/n 的/u 适用/v 范围/n 。

正确划分的 55 例,正确率为 55/71,即 78%。

但如果不予分词,而是按未分词前的字符串来查询,则查询出 3144 例,正确率只有 1.75%。可见,经分词软件加工过,核心库的分词标注信任度还是提高了许多。误差率会直接影响到词义标注的正确率,即使词义标注的其他环节都做得很好,词义标注的正确率也被限制在一个有限的范围。

2) 真实语料文本的差错

在对大规模的真实文本的处理中,错讹现象远不只是分词与标注的问题,其原因和类型是相当复杂多样的。如书写符号的正确与否,如“板登”(板凳)、“爱僧”(爱憎)、“反节命”(反革命)、“干于净净”(干干净净)、“蜜挑”(蜜桃)、“模梭两可”(模棱两可)、“青蒇高原”(青藏高原);词素的变更,如“美奂美轮”(美轮美奂)、“孤儿寡妇”(孤儿寡母);甚至不正常的空格,都会影响到词语的存在,如“不同 a”“按 p”“高锰酸 n 钾”等。

2000 万字“核心库”的分词单位是 16.38 万,排除掉字母串、符号串、数字串后的汉字词是 151,515条,从中再剔除出“错讹例”3621 条,相当部分就属于此。“错讹例”在词总数中所占比例只有 2.3%,它们的词频为 13,910 次,只占汉字词总频次 10,586,971 的 0.131%,如果从词频来看可以忽略不计,而从词语数来看,还是会有一定影响的。

3) 机器分词的强制性

机器分词看重相邻搭配的频率,只要经常结合在一起,表现出相当的凝固度,就会当作一个分词单位来处理。如“围成”“仅限于”“本市”“遥指”“中日”“攻下”“单靠”“这么回事”“(他)换好(了入殓的衣服)”“(一些—硫酸—)溅到(—了—腿上)”,都会当作词来看待。又如“德国—队、芬兰—队、日本—队、香港队”,前三个队是分,第四个“香港队”合;“劲歌—劲—曲”,这里“劲歌”合,“劲曲”分。因此,机器分词的结果一般叫“分词单位”,有的也叫“工程词”。[4]

3. “核心库”词表的构成

“核心库”的总汉字词 151,515 条。其中人名 27,557 条,地名 11,306 条,机构名 5963 条,数字词 3240 条,共计 48,066 条,占所有词语的 31.7%。具体见图 1。

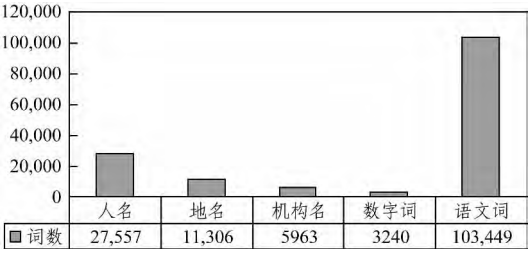

图 1　“核心库”汉字词的构成

图 1 显示了这样一些有价值的信息: (1) 四类专名占总词数的 31.7%,与“语文词”数量为 1∶2。(2) 语文词不仅数量多,而且使用频次也高。它是“人名”的 14 倍,“地名”的 7 倍,“机构名”的 20 倍,“数字词”的 7 倍。

再根据使用范围、来源等因素,又可将 103,449 条语文词分成“通用词”“行业词”“方言词”“古语词”四类。具体见图 2。

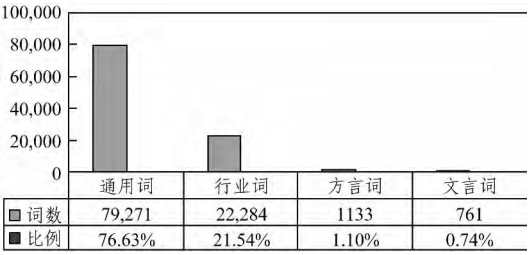

图 2　“核心库”语文词的构成

“通用词”占语文词中的大部分。其次是“行业词”。行业词即领域词,使用于某个具体的行业领域。它是通用词语的主要来源之一。还有“方言词”与“古词语”,前者重在地区差异,多用于某个地域,后者重在时代差异,保留着过去时代的特色。这两部分也是与通用词有着直接联系的,它们特色鲜明,数量较少,区别起来较为容易,影响通用词的程度较小。

通用词有一个重要特点,就是高频性。其实不管来源何处,只要具有了相当的频度与广度,也就具有了成为通用词的最重要条件。再辅之以稳定性的考察,能在较长时间内稳定地使用,确认为通用词也就不难了。

经过上面的分析后,现在可以对现代汉语断代词汇的构成提出这样的设想: 它有两个层面,下面一层是语言词。语言词的核心是通用词,这是整个现代汉语词汇系统的基本部分,居于中心位置。与其接壤并有着较强互渗力的是行业词、方言词、古词语,其中以行业词的影响最为重要。这三类词围绕在通用词的四周,中心部分与周边部分不是泾渭分明,而是有一个重合的宽阔地带。人名、地名、机构名、数字词则构成言语词层面,它们浮在语言词层面上。它们缺乏足够的稳定性,大都活跃在某个特定环境中。

(三) 词典词与真实语料词的差异及影响

词典词就是词典的词目。词典词与来自语言生活、来自真实语料的词是很不一样的。词典词重在继承,多源于前代辞书,它看重“考释性”,愈是人们不懂而需要查考的,愈是它的收录对象,而对“见字明义”的词则是不屑于收录的。另外,词目词总是落后于时代,词典词从来

都与语言生活实际用词隔有一定的距离。

人们转而开始重视从真实语言材料中提取词表,希冀达到对断代词汇整体面貌的了解。人们对此做过许多尝试,从专书研究时代起,就有通过选取代表作品、代表作家来窥探断代语言面貌的做法。到了计算机语料库时代,通过众多的作品、海量的语料来概括断代语言面貌更是成为一种普遍做法。从百万字级容量的语料库(产生了《现代汉语频率词典》[5]),直至千万级(产生了《现代汉语常用词词频词典》[6])、亿级(《中国语言生活状况报告》[7])容量的语料库。仅《人民日报》从创刊起至当下,容量就逾十亿字,一直成为人们非常关注的语料。而以平衡、通用、综合为特点的则以国家语委的"通用语料库"为代表。可新问题又来了,语言词与言语词共存,规范词与不规范词共存,各种语体色彩词共存,而且还在相当程度上存在着真实语料难以避免的杂芜、粗糙的情况。

因此,词典词与真实语料词也就有了许多不同的特点,有了书面语与口语词、正体词与异体词、语言词与言语词、通用词与领域词、稳定词与具体词等等方面的差异,这些差异都会在词的表现形式、使用频率、使用范围等统计项中顽强地表现出来。同时,从真实原始语料到可加工、可词义标注的过程,还会受到分词规则、分词方法、分词软件的影响。以上种种的词与分词的不同,都会在很大程度上影响到词义标注的效果与正确率。这都需要我们在建构词库时,在从事词义标注前,对词库的建构规模、性质、面对真实语料进行词义标注可能获得的效果做出估计,并采取相应对策。也就是说,词义自动标注的目标,或说是第一阶段的目标,只能是对典型的、稳定的、通用的、义项颗粒度较大的、属于语言体系内的多义词进行标注,而希冀对语言使用中所有的、特定场合的、临时的、语境的、属于言语层面的多义词都做出词义标注,这是不现实的。

基于这样的认识,我们建构了专门服务于词义标注用的义项库(Word Sense Base for Sense Tagging of Modern Chinese,缩写 WSB)。内含词语 8 万余条、义项 9 万余个。多义词 8 千余条,属多义词的义项有 1.3 万个。我们这里谈的词库是基于义项库来说的,关注的是"词"本身。它在整个 SCT 系统中居于基础位置,对整个 SCT 起着重要的支撑作用。概括地说,它是词义自动标注的来源、词义自动标注的对象、词义规则形成的依据。纳入首批考察、全面描写、检测的词语是:双音节、含 2 ~5 个义项、频次在 100 次以上、名动形三类词,共计 3771 条,义项 7861 个。

二 规则库的建库原则

(一)规则库的作用

规则库(SKB)是"多义词词义搭配知识库"(SCT)的另一个起着重要作用的核心库。规则库的作用可以这样来概括:它统摄语义库、义项库、语料库,这些知识库通过规则库的组织而发挥作用;它是实现词义标注工程目标的直接依据,对于任何一个多义词,规则定义的多寡有无、质量好坏直接决定标注结果;它集中体现 SCT 整个系统的意义与价值,是语言知识与工程实施的结晶体。

围绕着词义辨析,许多学派、许多学者都在不同研究领域做过探讨,都为 SKB 的研究提供了参考。SKB 的目标就是要为计算机描写多义词各个义项的区别性形式特征。要做到这一点,首先就要回答区别性词义形式特征有哪些,它们在具体语料中以怎样的形式存在,如何把握影响义项使用的语义制约,而邻近的特定共现词与共现词群为义项呈现提供了一个强制的语义环境。我们要做的就是通过对搭配对象或共现词的特征进行定位和约定,从而达到词义

自动识别的目的。其次,如何将这些义项形式特征进行描写,形成规则,并以可识别的方式提供给计算机,保证计算机可以准确、高效地利用这些规则,就成为一个关键。再次,如何组织利用其他资源,来为规则的提炼提供帮助。在这里我们特别看重语义库的利用,努力从语义搭配的角度,在一个统一的语义分类体系中以较小的义类为模块对具体词进行定位处理,在这个语义框架内约定其功能与位置,进而实现语义类别与语义特征的对接。[8]

资源库就好比是放在厨师面前的原料,没有它做不成菜肴,有了它也不等于是菜肴。要把原料做成可口菜肴还得经过烹调。SKB 就是这个烹调过程。简而言之就是要面对语料库,依托义项库,参照语义库和语法库,来抽取多义词的"义项区别性特征",再把这些"义项区别性特征"保存在规则库 SKB 中。只有经过了这个过程,菜肴才算正式完成。这个工作比较复杂,由于不同的词类有不同的规则构成,提取规则的模式与方法也有不同,故需要分别为名词、动词、形容词拟出不同的规则库,因此,三个词类课题小组也就在统一的规则库构建原则指导下具有了相对独立的探讨对象。完成规则库后,处理真实语料时后台资源库就隐身了,系统直接使用的是规则库、义项库、真实语料,把它们统一整合在执行软件中,就可以达到对真实语料自动进行词义标注的目的了。

(二) 影响规则库设计的因素

在汉语中,词汇搭配成为体现词语使用特征最主要的表现形式。可要把搭配习惯、规律以规则的形式提取、保留下来则是相当困难的。影响因素主要有:

1. 搭配内容庞杂与类聚的多样性

语言使用的灵活性决定了词汇搭配必然是多样、庞杂的。如"裁判"在 SCT 语料库中出现 503 次,以其前后 4 个词为界,所搭配使用的名词就有 883 个。把描写的对象放在词上,逐个描写必然耗时耗力,同时也难以应对语言使用实际状况。而根据结构主义的组合与聚合理论,利用《现代汉语分类词典》来观察其义类搭配规则可收到较好效果,用"语义库"来观察"裁判",发现光"司法"一个三级类就可以涵盖 129 个词,大大减轻了描写的困难。这只是一个初步的尝试。吴云芳等(2005)归纳了汉语中动词的宾语聚类的五种情况,发现有的类还会受到许多约束。[9] "语义库"的利用值得努力尝试,也是需要经过进一步探索的领域。

2. 语义、语法搭配位置的相对固定与绝对变动

有些词的区别性搭配总是出现在固定的位置,如"沐浴"有 3 个义项 "①洗澡。②比喻受润泽。③比喻沉浸在某种环境中。"一般情况下只有①可以直接跟"了"搭配。在这里"了"就具有了很好的鉴别作用,计算机也容易识别这样的规则。

可绝大多数情况下,搭配词的位置是不固定的。如,"编造"的"凭想象创造(故事)"义项跟"故事"搭配是比较稳定的,可它和"故事"之间的距离、位置却非常灵活。可以形成"编造故事""编造一个故事""那位作家用一生编造了许多扣人心弦的故事""这个故事是他编造的"等等。

3. 词类标注的利用与不足

现在分词与词性标注的正确率已达到相当高的程度,大多数情况下词性符号就可以直接利用来帮助识别多义词,如"补贴"有两个义项 "①贴补: ~家用 | ~粮价。②贴补的费用: 福利~副食~。"义项①是动词,义项②是名词。如下面的句子:

(11) 这/r 将/d 由/p 国家/n 予以/v 补贴/v", "生活/n 补贴/n 很/d 快/a 发到/v 灾区/n 人民/n 手/n 里/f 了/u。

计算机可以很容易判断出前者是①义,后者是②义。[10] 但在实际操作中词类标注并非都那么好用。如"补贴"在 SCT 语料库中出现了 914 次,标注为"vn"的有 234 次,占 25%,这时仅根据词性标注的结果就难以标注义项了,如:

(12) 暂停/v 对/p 职工/n 医疗/n 补贴/vn 和/c 住院/vn 医疗费/n 的/u 报销/vn。

4. 句法分析的重要及困难

句法分析对词义标注十分重要,特别是对 SCT 这类基于规则的系统。它需要语法分析重点解决这样两个问题:判断待标注词在句中的句法位置与搭配关系。目前的句法分析还难以达到实用程度,在这种情况下只能是在未充分依靠句法分析的语料中尽量保证计算机可以识别正确的匹配规则,又要有前瞻性地为以后利用句法分析框架留下接口。

(三) 规则库设计的基本原则

在为规则库设计标注内容、符号、空间时遵循着以下几个原则:

1. 利用多种语言资源尽量全面描写语义类聚,特别注重加强其覆盖力与区别度

"语义库"是我们自己研制的语义分类词典,它反映了词与词之间的语义远近关系,是 SKB 规则语义类聚的最主要依据之一。"词根"是汉语中固有的语义聚合资源,同根词以词根为连结点,在语义上往往有共同特征。如:

[标榜] ①提出某种好听的名义,加以宣扬。②(互相)吹嘘;夸耀。

义项①可以与"……主义"搭配,如"标榜社会主义、改良主义、爱国主义"等。以"主义"为词根的词很多,一般词典都难以收全,因此可以用"……主义"这一词根来概括。

再如:

[开发] ①以荒地、矿山、森林、水力等自然资源为对象进行劳动,以达到利用的目的;开拓。②发现或发掘人才、技术等供利用。③支付;分发。

义项①可以与地名搭配,如"开发新疆、开发浦东、开发湄公河"等,类似的结构、类似的词难以穷尽,这时可以用其共同词类标记"LOC"来概括。

利用聚类的优点是明显的,少量规则就可以涵盖大量具体的词语搭配,但聚类的同时不可避免会带来一些与其不相干的内容,如标"LOC"的词不一定都能与"开发"①搭配,以"主义"为词根的词也不一定都能与"标榜"搭配。语义库上下有五级类,上层类的概括力强,但带来的干扰因素也多,下层类"纯净",但概括力受限。这时可以采用先用上层类以求覆盖、后用下层类以求剔除的方法。

2. 语法框架与语法位置相结合

对句法关系的语法框架是需要的,但在目前尚不能得到充分利用的情况下,采用在语法框架中容纳语义规则的做法,即规则既包含了对搭配内容的规定也包含着对搭配所充当的语法属性的规定。仍以"开发"为例,它的搭配规则包括:义项编号为 16201,客事为 LOC。其中,义项编号是"开发"①在义项库中的编号,"客事"表示语法关系和属性,"LOC"表示搭配的语义内容。

3. 按不同词类分别设计规则框架

动词、名词、形容词三种词类在句法特点与搭配上都有很大的不同,因此分别为其"量身定做",设计各自的搭配框架,单独制定规则,也就很有必要了。

4. 充分利用多义词义项的不平衡性来制订排除性规则

排除性规则是指不正面描写一个义项可以与哪些成分搭配,而是列出不能与之搭配的规则。如"安排"有两个义项 "①有条理、分先后地处理(事物);安置(人员)。~工作丨~生活丨~他当统计员。②规划;改造。重新~家乡的山河。"义项②不仅低频,语境也比较简单,在 SCT 的 2.5 亿字的语料库中只出现过"安排……山河"一种用法,如"林县人民多壮志,誓把山河重安排','誓把下马山山河重安排','为了要把山河重安排"。这说明它的出现环境有着特定性,也缺乏扩大使用的活力,而义项①的搭配则极为丰富,要详细描写相当困难。因此在给义项①的规则描写时只要在客事位置排除"山河"这条规则就可以了,其效果要好于详细列出这个义项可以与哪些词搭配。

[参考文献]

[1] 李建国.《现代汉语词典》与词汇规范[A].《现代汉语词典》学术研讨会论文集[C].北京:商务印书馆,1996.

[2] 苏新春,顾江萍.人、机分词差异及规范词典的收词依据[J].辞书研究,2001,(2).

[3] 国家语委现代汉语语料库介绍[R].教育部"中国语言文字网",http://www.china-language.gov.cn/25/2007_6_20/1_25_340_0_1182320940531.html.

[4] 卞成林.基于信息处理的汉语工程词研究[J].广西民族学院学报(哲学社会科学版),1999,(1).

[5] 北京语言学院语言教学研究所.现代汉语频率词典[Z].北京:北京语言学院出版社,1986.

[6] 刘　源.现代汉语常用词词频词典[Z].北京:中国宇航出版社,1990.

[7] 国家语言资源监测与研究中心.中国语言生活状况报告(下编)[R],北京:商务印书馆,2006.

[8] 苏新春.现代汉语分类词典[Z].北京:商务印书馆,2013.

[9] 吴云芳,段慧明,俞士汶.动词对宾语的语义选择限制[J].语言文字应用,2005,(2).

[10] 王　惠.现代汉语名词词义组合分析[M].北京:北京大学出版社,2004:220.

广东省中国语言学会举行年会

广东省中国语言学会于 2013 年 12 月 7~9 日在广州中山大学举行 2012-2013 学术年会,本届年会参加者共 94 人,提交论文 87 篇,包括语言学研究的方方面面,会议气氛和谐团结、充满朝气。报告精彩纷呈,讨论热烈活跃,彰显开放、务实、多元的学术特色。中山大学党委书记郑德涛出席开幕式并致欢迎词,会长邵敬敏致开幕词,副会长张玉金致闭幕词。会议还特邀黄正德(美国)、柯里斯(法国)、徐杰(澳门)以及石锋、沈阳等海内外著名学者发表演讲。学会还编撰了第 15 期"简讯",对两年来的学会工作以及各高校的语言学研究动态进行全面总结回顾。

年会期间,举行了理事会、常务理事会以及会员大会,选举产生了第八届理事会。为促进新陈代谢,让优秀的中青年学者进入领导班子,本届理事会除 24 名理事续任外,新当选的理事有 19 名,新增选常务理事 4 名。邵敬敏继续担任会长,李炜出任副会长兼秘书长,林伦伦、周小兵、张玉金以及屈哨兵继续担任副会长。会议还聘请詹伯慧担任名誉会长,唐钰明为学术委员会主任,周娟为副秘书长。

(周娟)