

动宾组合的自动获取与标注

陈小荷

北京语言文化大学语言信息处理研究所

摘要：词语搭配（特别是抽象词语的搭配）是汉语自动句法分析的一个重要知识源。我们从 50 万字词性标注语料中自动获取动宾组合实例，并且将未经校对的搭配数据用于动宾结构的自动标注，标注正确率和召回率分别为 74.7%和 76%左右。结果表明，词语搭配数据对解决真实文本中动宾组合竞争问题十分有效。

关键词：动宾组合 组合竞争

Automatic Acquisition and Bracketing of Verb-Object Collocations

Chen Xiaohu

Institute of Language Information Processing, Beijing Language & Culture University

Abstract Collocations play an important role in parsing particularly when they comprise abstract words. 21628 candidate verb-object collocations were automatically obtained from a corpus with 200000 words. We then bracketed the verb-object phrases in the corpus by using the data without any correction and the precision and recall are about 74.7% and 76% respectively. It shows that the knowledge of collocations is very effective in resolving the verb-object combination competition in real texts.

Keywords VERB-OBJECT COLLOCATION, COMBINATION COMPETITION

1. 引言

句法分析（parsing）是通向自然语言理解的关键步骤。实践证明，句法分析中仅仅使用语法知识是不够的，语义知识和词语搭配知识是另外两种基本的知识源。汉语缺乏形态，因此在汉语的句法分析中，后两种知识源显得更为重要。

语义研究是整个语言学研究中比较薄弱的环节，如何利用语义知识来支持句法分析，我们还缺乏成功的经验。一个朴素的想法是，既然单纯利用语法知识会产生太多的歧义，那么在句法分析中加上一些语义限制，也许可以排除或减少歧义[1]，[2]。但这样做的前提是，有一个合适的语义分类体系，可以轻易地将每一个（或绝大多数）实词归入这个或那个语义类别，以便清楚地表述我们心目中的那些语义限制条件。例如，把“杨树”、“柳树”等归入“树”，动词“栽”的受事是“树”这个语义类别（或许还可以概括为“植物”），从

而可以解决动词“栽”的动宾组合问题。但这样简单的例子实在不多。对表示具体事物的实词进行语义分类比较容易，对表示抽象事物的实词进行语义分类则十分困难。^① 例如，要问“困难”、“水平”之类的抽象名词属于什么语义类，就很难有一致的回答，甚至也不清楚是否存在这样的语义类。“困难”、“矛盾”、“问题”等都可以做动词“解决”的宾语，它们似乎有某种共同的语义特征；“水平”、“觉悟”、“战斗力”等都可以做动词“提高”的宾语，能否概括为某个语义类或指明其共同的语义特征以说明这些动宾组合的语义限制条件，使得符合条件的词语都可以做“提高”的宾语而不符合条件的词语都不能做“提高”的宾语呢？恐怕很难做到。即使可以做到，也首先需要把这些组合实例找出来。

我们正在做一个现代汉语自动句法分析的项目，目标是处理真实文本。基本设想是，建立一个为句法分析服务的最小语义体系^②，以描写表示具体事物的词语的语义组合限制；建立一个关于实词（特别是抽象词语）搭配的电子词典，为句法分析提供词语搭配知识。已出版的几部实词搭配词典[3]，[4]，分别有几十万条实词组合实例，有人用这种词典进行语义联想网络的研究或语义自动聚类研究[5]，[6]，取得了一些令人满意的初步结果。但是直接用这种搭配词典来支持句法分析，存在两个主要问题：第一，表示抽象事物的词语的搭配虽有典型实例，但不充分，难以处理较大规模的真实文本；第二，缺乏搭配数据，难以有效地处理组合竞争问题。因此，需要从大规模真实文本中来自动获取这种组合实例及其搭配数据。

对现代汉语真实文本进行完全的句法分析，目前还不现实；我们的基本策略是先进行多趟不完全分析以取得经验。在多趟分析中，先做主谓、述宾和介宾这几种结构跨度较大的短语的分析，然后抽取出剩余的短语做后续分析。本文讨论动宾组合实例及其搭配数据的获取，以及如何用这些知识来自动标注真实文本中的动宾结构。

2. 语料说明

我们使用的是已经标注词性的科技、政治新闻类的报纸语料，约 50 万字，20 万词次。这两类语料基本属于书面语体，生活方面的词汇较少，表示抽象事物的词语较多且频率高，这一点正符合我们的需要。

这些语料是北京语言文化大学的“现代汉语研究语料库”[7]的一部分，词性标记集包括 112 个标记。^③ 该标记集的一个重要特点是提供了较多的句法信息，例如动词有以下 16 种标记：

va	助动词	vc	V 作补	vf	形式动词
vg	V 不带宾	vga	V 带形宾	vgb	V 作宾
vgd	V 带双宾	vgj	V 带兼语宾	vgn	V 带体宾
vgp	V 作 NP 偏	vgs	V 带小句宾	vgv	V 带动宾
vgx	V 作 NP 正	vgz	V 作主	vi	系动词
vv	“来去”+VP				

其中很多是“动态的功能属性”，如 vgn 和 vg 不是通常的及物动词与不及物动词的分别，而是指该动词在当前文本中是否带了宾语。众所周知，汉语词类是多功能的，动词可以

充当许多中句法成分，但具体到每一个动词有哪些功能，在具体的语境里实现的是哪一个功能，需要作具体分析。这里 *vgz*、*vgb*、*vgx*、*vgp* 分别表示动词在具体的语境里充当主语、宾语、名词短语里的中心语和定语。

名词有以下 7 种标记：

nf 姓氏 *ng* 普通名词 *ngl* 离合名词 *npf* 人名
npr 其它专名 *nps* 地点专名 *npu* 机构名

总的来说，用这个标记集标注的语料更有利于句法分析，其中，动词的“动态的功能属性”的标注实际上已经是句法分析的一部分。于是，在获取动宾组合时，就可以排除 *vgz*、*vgb*、*vgx*、*vgp* 作支配成分的可能。我们只把 *vgn*、*vg* 这两种标记的动词作为获取动宾组合的驱动点，这是因为：

第一，我们只对实现了“带体词作宾语”这一功能的动词感兴趣，标记这一功能的是 *vgn*；

第二，*vi* 标记的主要是动词“是”，“是”虽然也常常带体词性宾语，但它的组合能力太强，获取这种组合实例没有什么实际意义；

第三，*vg* 标记的是未实现“带宾语”这一功能的动词，但有时在语料中事实上带了宾语。这可能是由于动词离宾语的核心名词距离较远，或者是因为右邻是 *vc*，如：

邓小平/*npf* 同志/*ng* 作/*vg* 出/*vc* 了/*utl* “/ 科学/*ng* 技术/*ng* 是/*vi* 第/*maf* —
/ *mx* 生产力/*ng* ” / ” 的 / *usde* 英明 / *a* 论断 / *ng* ; / ;

把握/*vg* 住 / *vc* 改革 / *vg* 、 / 、 发展 / *vg* 和 / *c* 稳定 / *a* 的 / *usde* 大局 / *ng* , / ,

不过，*vg* 有可能是从来不带宾语的动词。为了减少统计误差，我们利用北京大学开发的《现代汉语语法信息词典》(电子版) [8]，把从来不带体词性宾语的 *vg* 过滤掉。

名词中我们只关注普通名词 *ng*，但是把动词中的 *vgx* 标记也考虑在内，这是因为 *vgx* 标记有两种可能，一是真正的动词，二是兼属动词的名词，后者如：

小平 / *npf* 同志 / *ng* 的 / *usde* 指导 / *vgp* 思想 / *ng* 和 / *c* 泽民 / *npf* 同志 / *ng* 的 / *usde*
重要 / *a* 指示 / *vgx*

更重要的原因是，即使 *vgx* 是真正的动词，当它又充当另一动词的宾语时，这种动宾组合往往也是一种习惯性的搭配，值得关注，例如：

世界 / *ng* 各 / *ra* 国 / *ng* 都 / *dr* 加强 / *Hvgn* 了 / *utl* 对 / *pg* 科学 / *ng* 技术 / *ng* 的 / *usde*
规划 / *vgx* 和 / *c* 控制 / *vgx* , / ,

世界 / *ng* 范围 / *ng* 内 / *f* 基础 / *ng* 学科 / *ng* 结构 / *ng* 发生 / *Hvgn* 明显 / *a* 的 / *usde*
变化 / *vgx* 。 / 。

3. 动宾组合的自动获取

我们采用统计方法来获取动宾组合实例。具体地说，统计每一个动词(标记为 *vgn* 或 *vg*)跟出现在它后面的名词(标记为 *ng* 或 *vgx*)的同现次数，根据同现次数计算它们的关联(association)程度，选取关联程度较高的组合进入搭配词典。^④

关联程度的计算有多种方法，这里先用互信息作为对动词和名词宾语的关联程度的估计。一般是在确定了动宾关系之后才作这种计算[9]，但我们现在的目标是先发现可能的动

宾组合，因此实际上是计算“可能的动宾组合”中动词跟名词的互信息，计算公式如下：

$$I(v;n) = \log \frac{P(n \text{ 出现在 } v \text{ 之后})}{P(n)P(v)}$$

计算互信息时，所谓“同现”通常是指紧挨着出现。但是，动宾结构的平均跨度较大，仅仅统计相邻同现次数只能找出“发展经济”之类的组合，不能找出“作…论断”这样的离散同现的组合。我们设定的观察窗口是句子或分句，其长度不固定，一个动词和一个名词只要是在同一个句子或小句里同现，且动词在前，名词在后，就算是一个可能的动宾组合实例。单纯按上述方法来统计动词和名词宾语的同现次数会有较大的误差，这些误差主要体现在两个方面：第一，动词（特别是标记为 *vg* 的动词）后面出现介词或助词“的”时，实际上已经不可能带宾语，应予排除：^⑤ 第二，若干个名词连续出现（中间没有任何别的词类）时，最后一个名词作为宾语中心词的可能性最大，例如：

维护/*vgn* 法律/*ng* 权威/*ng*

未/*dr* 作/*vgn* 资信/*ng* 调查/*vgx*

因此对这种连续出现的名词，我们先把它捆绑起来，假定最后一个名词是中心词；这样上述两例就只统计“维护…权威”、“作…调查”的同现次数，从而较大幅度地排除了“维护…法律”、“作…资信”这样的伪组合。^⑥ 当然会有少数例外，如：

关于/*pg* 落实/*vgn* 知识分子/*ng* 政策/*ng* 问题/*ng*

后面三个名词本不应捆绑。但是我们容忍这一点误差，相信只要有相当规模的语料，能够找到“落实…政策”这样的组合实例。

另外，名词后面若出现介词，该名词一般也不会做前面动词的宾语，应予排除，例如：

预测/*vg* 到/*vc* 信息/*ng* 对/*pg* 未来/*t* 社会/*ng* 发展/*vgx* 的/*usde* 重要/*a* 作用/*ng*

充分/*az* 强调/*Hvgn* 了/*utl* 科学/*ng* 技术/*ng* 对/*pg* 经济/*ng* 发展/*vgx* 和/*c* 社会/*ng* 进步/*vgx* 的/*usde* 重要/*a* 作用/*ng* , /,

将明显不能作为述语的动词以及不能作为宾语的名词排除，并且将连续出现的名词捆绑之后，如果句子（或分句）中只剩下一个动词和一个名词，它们作为动宾组合的概率是较高的；如果此时句子中有若干个动词和名词，到底哪个动词跟哪个名词组合，没有更多的信息则难以确定。这时有两种选择：一是只统计前一种情况，这样得到的动宾组合可信度高，但需要有大规模的词性标注语料；二是对两种情况加以区分，虽然获取结果的可信度会有所降低，但能够充分利用宝贵的语料。我们选择后一种办法，看当前动词后面有几个候选的名词宾语，如果只有一个，则同现次数为一；如果有几个，则平分一个同现次数。例如：

使/*Hvgj* 之/*rs* 成为/*vgn* 各/*ra* 级/*qns* 政府/*ng* 部门/*ng* 制定/*vgn* 政策/*ng* 和
/c 计划/*ng* 的/*usde* 依据/*ng* , /,

其中“成为”后有四个名词，则“成为…部门”、“成为…政策”、“成为…计划”和“成为…依据”各算 1/4 次。“制定”后有三个名词，计算方法同上，每个候选组合各算 1/3 次。^⑦

我们用上述方法从 50 万字语料中获取了 21628 个可能的动宾组合，其中出现次数大于或等于 1.0 次的组合有 12674 个。前 500 个互信息最高（12.21~11.11）的组合，正确率为 70%，其中含抽象名词的组合的正确率约为 75%，这两个结果都不甚理想。

第二次我们按同现次数排序，也检查了前 500 个最高频（92.82~4.41 次）的组合，正确

率为 87%，其中含抽象名词的组合同正确率为 93%。通过分析数据，发现按互信息排序时，前 500 个组合平均只出现 0.96 次，显然是由于同现次数太少而影响了这些互信息高的组合的正确率。

以“关系”为宾语的组合实例，《现代汉语实词搭配词典》[4]列举了 19 个，而我们仅从 50 万字语料中就获取了 61 个。其中相同的有 11 个（只列动词，下同）：

保持 断绝 发生 分析 改变 改善 利用 了解 调整 维护 影响

《现代汉语实词搭配词典》有，而我们未获取的是 8 个：

存在 产生 挑拨 离间 割断 搞好 研究 弄清

我们已获取，而《现代汉语实词搭配词典》未收的是 50 个：

安排 把握 变更 表现为 查清 阐明 冲淡 处理 促进 调节 发展

巩固 固定 规范 恢复 加强 建立 讲明 解决好 界定 扩大 拉

理解 理顺 密切 明确 明晰 评价 破坏 确保 认识 深化 盛赞

谈到 推进 脱离 拓宽 拓展 维持 稳定 削弱 协调 形成 修补

修复 有着 掌握 珍惜 重视 主导

虽然仅一个词例不能全面地比较两种来源的动宾组合，但也可以看出一些端倪。人用的搭配词典由于篇幅限制、搜集材料的方法等各种原因，只能列出若干典型实例，象“拓宽…关系”（“进一步拓宽互利交流和合作关系”）、“主导…关系”（“经济优先已成为主导国家关系的重要因素”）这样的非典型实例不易想到^⑥，而“理顺…关系”、“建立…关系”这样的典型实例也不能尽数列举。

4. 用于动宾标注

为了检验搭配数据的价值，我们将这些数据用于动宾结构的自动标注。标注对象仍然是这 50 万字语料，而且对所获取的动宾组合实例没有作任何人工校对，因此相当于开放测试。基本的标注算法如下：

（1）扫描句子中的动词、名词（按上一节的方法进行了筛选和名词短语捆绑），将候选动宾组合存入一个数组；

（2）如果数组为空，转（5）；

（3）取数组中互信息最高的候选组合作为输出；

（4）从数组中删去该组合，以及跟该组合相冲突的所有组合，转（2）；

（5）结束。

例如，对于句子：

科学/ng 技术/ng 进步/vgx 已经/dr 成为/vgn 一/mx 个/qng 国家/ng 发展
/vgn 经济/ng 、/、 增强/vgn 综合/b 国力/ng 的/usde 关键/ng 。/。

扫描到候选动宾组合：

成为…国家	成为…经济	成为…国力	成为…关键
	发展…经济	发展…国力	发展…关键
		增强…国力	增强…关键

其中“发展…经济”的互信息最高，输出之，并且将“发展…经济”、“发展…国力”、“发展…关键”、“成为…经济”都删除。

在剩下的候选组合中，“增强…国力”的互信息最高，输出之，并且将“增强…国力”、“增强…关键”、“成为…国力”都删除。

最后剩下的两个候选组合中，“成为…关键”的互信息最高，输出之，并且将“成为…关键”、“成为…国家”都删除。此时数组已空，结束。输出的结果如下（用方括号表示动宾结构的左右边界，可嵌套，方括号内侧的数字表示输出的顺序而非层次顺序）：

科学/ng 技术/ng 进步/vgx 已经/dr [3 成为/vgn 一/mx 个/qng 国家/ng [1 发展/vgn 经济/ng1] 、 /、 [2 增强/vgn 综合/b 国力/ng2] 的/usde 关键/ng3] 。 /。

此外还有一些具体的技术处理，例如一个句子（或分句）中出现几个相同的动词或几个相同的名词，按就近原则组合；连续两个动词，若能根据它们跟某个名词的互信息的比值确定为联合结构，则扩大动宾组合的左边界，等等。

我们对 29 个文本的标注结果进行了人工校对，标注正确的动宾结构 1501 例，标注错误的 509 例，正确率为 74.7%；应标注的动宾结构 1976 例，召回率为 76%。校对时发现一个明显的倾向：含有抽象名词的动宾组合，其标注的正确率和召回率都比较高，这一点是我们所乐见的。至于具体词语的组合，本来就不指望用词语搭配来完满地解决。

以上算法是把互信息作为处理动宾组合竞争的一种依据。以“成为”为例，无论从词性还是语义上看，“国家”、“经济”、“国力”、“关键”都有可能做它的宾语，形成了一个动词跟若干个候选名词宾语的组合竞争；另一方面，一个名词（如“关键”）也可能跟若干个候选动词述语（“成为”、“发展”、“增强”）形成组合竞争，就这个例子而言，也许能从句法或语义方面找到选择标准。但是用我们的算法无须制定复杂的规则，而且由于是基于词语搭配的优选，因此当动宾结构内部含有某些语法错误时也能表现出较强的鲁棒性。在应标注的 1976 例中，有 525 例是经筛选、捆绑后没有动宾组合竞争的，姑且认为这 525 例应该全部标注正确，^⑨ 则标注程序处理动宾组合竞争的正确率为 $(1501-525) / (1501-525+509) = 65.7\%$ ，召回率为 $(1501-525) / (1976-525) = 67.3\%$ 。

本文报道了从一个规模不大的语料库自动获取动宾组合实例并将搭配数据用于真实文本的动宾结构标注的实验。实验结果表明，从语料库获取的组合实例能更充分地反映语言面貌，而搭配数据可以比较有效地解决组合竞争问题。

我们使用的是已经词性标注和人工校对的语料，而且标记中含有较丰富的动态语法信息，因此动宾组合的获取和标注对特定的词性标记集有较强的依赖性。显然，这种已加工语料是昂贵的和不可多得的，需要进一步研究如何从原始语料或简单词性标注语料中获得组合实例及其搭配数据。另一方面，含抽象词语的动宾组合在书面语体的语料中一般都有较高的出现概率，因此，如果语料库规模扩展到千万词级，估计就可以把这种组合基本收录进来。

参考文献

- [1] 黄曾阳，HNC（概念层次网络）理论，清华大学出版社，1998年，北京。
- [2] 陈小荷，一个面向工程的语义分析体系，《语言文字应用》，1998年第2期。
- [3] 倪文杰、张卫国等，现代汉语辞海，人民中国出版社，1994年，北京。

- [4] 张寿康、林杏光, 现代汉语实词搭配词典, 商务印书馆, 1996年, 北京。
- [5] 苑春法等, 汉语语义关联网的研究, 《语言工程》, 清华大学出版社, 1997年, 北京。
- [6] 李涓子等, 基于组合实例的双向优化聚类, 《语言工程》, 清华大学出版社, 1997年, 北京。
- [7] 孙宏林等, “现代汉语研究语料库系统”概述, 《第五届国际汉语教学讨论会论文集》, 北京大学出版社, 1997年, 北京。
- [8] 俞士汶等, 现代汉语语法信息词典详解, 清华大学出版社, 1998年, 北京。
- [9] 翁富良、王野翔, 计算语言学导论, 中国社会科学出版社, 1998年, 北京。

附注

- ① 文献[1]对此有完全相反的观点, 认为抽象词语的语义描写反而比具体词语容易。(第41页)
- ② 语义描写可深可浅。“最小语义体系”是较浅的语义描写手段, 不包括那些跟句法分析无关或关系不大的语义描写。
- ③ 文献[7]对该语料库和标记集有详细介绍。我们使用的是该语料库的旧标记集, 但跟新旧标记是一一对应的。
- ④ 关联程度较高不一定是真正合法的组合, 所以进入搭配词典之前需要人工校对。
- ⑤ “动词+‘的’+名词”序列中, 名词可能是动词的逻辑宾语, 如“领导/vg 者/kn 自己/rs 看好/vg 的/usdc 项目/ng”。如果能准确地判断出这种关系, 获取动词的语义配价时也可加以利用。
- ⑥ “维护…法律”在别的上下文中可能是一个正确的动宾组合, 但我们还希望尽可能准确地统计动宾组合的出现次数。
- ⑦ 如果计算机能有把握确定“政策和计划”是联合结构, 就可以进一步减少候选组合的数量, 例如, 本例中“制订”的可能的名词宾语就只有“计划”和“依据”, “制订…计划”和“制订…依据”都算出现了1/2次。
- ⑧ 从语义上看, 非典型实例不一定是恰当的组合, 例如“拓宽…关系”似乎就有点问题, 但这也是处理真实文本时不能回避的。词语搭配知识的利用, 在这个问题上正好可以增强一点分析器的鲁棒性。
- ⑨ 如前所述, 简单地将连续名词捆绑起来会有一些错误, 这525例并非完全标注正确。因此, 后面在处理动宾组合竞争的正确率和召回率的估计是较为保守的。