

# 现代汉语动名语串结构关系判定<sup>1</sup>

邱立坤

北京城市学院人工智能研究所, 北京, 100083

---

## 提要:

本文从大规模标注语料库中获取实例,使用电子版现代汉语语法词典和现代汉语语义词典,探讨现代汉语动名语串结构关系判定的问题。本文考察了分别引入动词、名词次范畴,语法词典中的动词直接修饰名词与名词直接受动词修饰两种属性,语义词典中动词、名词的语义属性,来对动名语串结构关系进行判定的效果。

**关键词:** 动名语串 结构关系 自动分析 现代汉语 标注语料库 电子词典

---

## 1 本文所用资源

本文使用了北京大学计算语言学研究所 PFR 语料库、《现代汉语语法信息词典》电子版局部、《现代汉语语义词典》电子版局部。

所用大规模标注语料库为人民日报社新闻信息中心、北京大学计算语言学研究所、富士通研究开发有限公司三方于 2001 年 4 月 1 日免费公布的“PFR 人民日报标注语料库(版本 1.0, 下面简称 PFR 语料库)”(文本内容为 1998 年 1 月份人民日报),计有 110 余万词。该语料库在标注过程中所采用的标记集<sup>2</sup>中作出了名动词(Vn)与动词(v)的区分。

本文所用的《现代汉语语义词典》是中科院计算所与北京大学计算语言研究所联合开发的一个汉英机器翻译系统中的一个重要组成部分。该词典已对 4 万余条汉语常用实词的语义分类及语义搭逐一作了描述。本文在分析过程中主要使用了该词典关于动词主体语义类、客体语义类,名词语义类的描述信息。其中,动词库词条 10788 项,名词库词条 27828 项。

本文所用的《现代汉语语法信息词典》为北京大学于 1999 年 7 月扩充后的版本,词典收词已达 7.3 万余条,并且完成了归类。本文在分析过程中主要使用了该词典关于动词“后名”、“外内”<sup>3</sup>等属性,名词“前动”属性的描述信息。

---

<sup>1</sup> 本文使用了北京大学计算语言学研究所提供的大规模标注语料库、语法信息词典和语义词典,特此致谢。

<sup>2</sup> 请参见《规范与手册·附表》。

<sup>3</sup> “外内”属性用于描述动词是否及物,不及物动词填“内”,及物动词不填。

## 2 识别动名语串结构关系的价值及相关研究

### 2.1 标注语料库对动词直接修饰名词现象的处理及其可能的价值

动词与名词搭配形成的动名语串可能为述宾结构或者定中结构，其中，定中结构的动名语串即属于本文所讲的动词直接修饰名词现象，在本文中两者指同一概念。根据北大计算语言学研究所的词类标记集，在词性标注过程中，需要将名动词与一般动词区分开。在经过软件自动切分与标注之后，我们发现：一方面，对于名动词与动词的区分，目前还没有有效的处理方式，从而使得这两个类的判别成为人工校对过程中的主要工作之一；另一方面，则使我们想到，如果在词性标注阶段能够多解决一些问题，实际上也就为句法分析提供了更多的信息。

从寻找谓语中心语的角度出发，做出名动词与一般动词的区分实际上也就是在进行排除某些动词成为谓语中心语的工作，而这显然是属于句法分析层面的问题。从句法分析的角度出发，动词是控制句子结构的关键一环，找出了谓语中心语，就等于完成了句法分析工作的一半。从这个意义上说，在词性标注过程（而不是句法分析过程）中，做出关于动词与名动词的区分似乎显得有些不必要，因为句法分析的前提就应当是词性标注。倘若词性标注阶段尝试无法较好地得到解决的问题，而句法分析阶段又以此为前提进行，那么势必造成两个阶段的连锁不良反应。

但是，北大计算语言学研究所机器自动加工辅以人工校对的方式所建立的大规模标注语料库为我们提供了大规模的动词直接修饰名词的实例，这使得我们可以设想通过在语料库中寻找规则，来实现在词性标注阶段<sup>4</sup>基本解决名动词与动词的区分问题的目标，从而为句法分析阶段寻找谓语中心语的工作奠定基础。这也正是标注语料库的用途与价值所在。

### 2.2 本文所面临的问题及将予以讨论的问题

动词除了成为名动词及副动词外，仍然可能不是谓语中心语。事实上，在谓语中心语的判定过程中，我们面临的正是以下几个难题：

- (1) 动词在何时成为名动词，包括三种情况：直接受名词修饰与直接修饰名词、直接受“的”字结构修饰成为体词性短语的中心语、在形式动词或其它准谓宾动词及“有”之后。
- (2) 以谓词或谓词结构的身份成为“的”字结构的中心语。
- (3) 以谓词或谓词结构的身份出现在介词结构中。

<sup>4</sup>笔者以为词法分析与句法分析并非是截然分开的，词法分析做得细一些，句法分析就容易一些。在此做出 Vn 与 N 的区分，实际上就是在进行句法分析的工作。因为一旦确定某词为 Vn，则 Vn 与其后的名词或者名词词组一定会是定中关系，从而排除了述宾关系的可能，也就排除了结构整体成为谓语中心语的可能。

以上几个问题没有解决,而仅仅依靠所谓的完句成分<sup>5</sup>及其它一些语法标记来判定谓语中心语,自然不可能得到理想的效果。

本文因篇幅与时间所限,不可能对以上几个问题一一展开,只是打算对(1)中的一个部分,即动词直接修饰名词(此类动词即具有“后名”属性)的现象进行考察,将其同与之同形的述宾关系的动名语串区分开,以图部分地解决名动词与动词的区分问题(同时也就可以确定某些动词不是谓语中心语)。

### 2.3 相关研究:应用动词次范畴和名词次范畴<sup>6</sup>的组合

(傅承德,1993)讨论了通过动词次范畴与名词次范畴相结合,并辅以词汇驱动规则来实现动名语串动宾与定中两种句法结构的自动识别的问题。它通过划分次范畴(动词分为及物动词与不及物动词,名词分为抽象名词与非抽象名词,单音节名词、双音节名词),然后将动词与名词的次范畴综合起来考虑,找出在搭配中结构关系不确定的情况,并采取词汇驱动规则来解决。由于该文仅考察了94个动词与106个名词,所以只能算是一项实验性的研究。

(李晋霞,2002)基于句法的角度,提出了面向计算机的动名语串结构关系识别策略。大致如下:(1)首先分出唯定中动词和唯动宾动词(唯定中动词是只能作定语而不能带宾语的动词,唯动宾动词则是不能作定语只能带宾语的动词),(2)其它的动词称为定中、动宾两可动词,对于由这一类动词构成的动名语串,主要通过给出高频名词表(包括157个名词)的方式解决结构关系识别的问题。

中文信息处理学界在识别名词短语时,通常采用以统计为主的方法。从这几年来的一些研究实践来看,自动处理效果并不是很理想。主要的研究工作包括:

(1)李文捷等人利用边界分布信息构造概率模型而进行的MNP自动识别实验,其开放测试的识别正确率达到了71.3%(在30篇新闻报道语料中)。<sup>7</sup>

(2) Angel S. T. Tse, 等人利用统计和规则相结合的方法,构造了“的”字名词短语自动分析器。实验结果为:识别正确率为75%,召回率为90%(在15篇汉语文本中)。<sup>8</sup>

(3)(周强、孙茂松、黄昌宁,2000)通过对包含5573个汉语句子的语料文本中的最长名词短语的分布特点的统计分析,提出了两种有效的汉语最长名词短语自动识别算法:基于边界分布概率的识别算法和基于内部结构组合的识别算法。实验结果显示,后者的识别正确率和召回率分别达到了85.4%和82.3%。

### 3 动名语串结构关系的判定

<sup>5</sup> 见杨成凯《广义谓词性宾语的类型研究》。

<sup>6</sup> 引入次范畴进行分析实际上也就是引入了语义。在本文中,后面所讲的引入语义与此处的使用次范畴相比,在语义的分类与描述上要更为细致和深入。

<sup>7</sup> 转引自周强、孙茂松、黄昌宁(2000)。

<sup>8</sup> 同上。

经过中文信息处理学界多年的辛勤工作，基础知识库的建设已经取得了巨大的进展，北大的《现代汉语语法信息词典》和董振东的《知网》在网络上都有版本可供免费下载以供研究之用。为了检验词汇主义路线的有效性，本文利用北大计算语言学研究所的《现代汉语语法信息词典》和《现代汉语语义词典》中部分信息进行了动名语串结构关系判断的实验。

### 3.1 应用语法词典中动词的后名属性与名词的前动属性

首先，我们利用《现代汉语语法信息词典》关于动词的后名属性与名词的前动属性的描写来进行实验<sup>9</sup>。我们在 PFR 语料库中抽取了 10000 例“v+n”语串，去掉重复的部分之后得到 7339 例，去掉动词为单音节的实例<sup>10</sup>后得到 5303 例。本文对 5303 例语串分析的结果如表一（在表一中，我们根据语串的动词部分与名词部分是否具有相应属性对其进行归类）：

表一

动词是否具有后名属性	是	是	否	否
名词是否具有前动属性	是	否	是	否
语串数量	479	1105	935	2784

我们的判断规则为：如果动词具有后名属性，同时名词具有前动属性，那么动名语串的结构关系判定为定中关系。由上表可知，由具有后名属性的动词与具有前动属性的名词组成非定中结构动名语串的实例共有 479 条<sup>11</sup>，根据判定规则，这 479 例判断错误：应为定中关系，而事实上却是述宾关系；其余判断正确。因此，判断正确率为  $(7339-479)/7339=93.5\%$ 。

动词的后名属性和名词的前动属性本来是用于描述动词直接修饰名词这一现象的，也就是说专门针对定中式动名短语的，照说如果用定中式动名短语来做实验，正确率应该很高。所以，我们重点选取了述宾式动名短语来实验，正确率竟相当之高。假设采用同样的步骤分析定中结构动名语串的正确率为 100%，则上述结果是非常有

<sup>9</sup>我们假定词典的描述全部正确，然后再利用词典所描述的信息进行判定。当然，这种假设是出于处理的需要而作出的，事实上词典的描述是需要不断改进的，尤其是在面对大量真实语料的时候，词典收词的覆盖面不可能达到 100%，属性描述也不可能尽善尽美。所以，语料库的建立、发展与词典的建立、完善是一个互动的过程。

<sup>10</sup>在 7339 例语串中，名词为单音节的数量为零；而动词为单音节的动名语串基本没有构成定中结构的例子，所以将其去掉不会影响我们对定中结构动名语串数量的统计。

<sup>11</sup>与此数据相对应，由符合此条件的单音节动词与双音节的名词组成的语串数量为 290。

效的, 可惜的是, 用定中式动名短语来做实验时, 正确率竟不到 60%。之所以如此, 一方面是因为词库整体规模不够大, 碰到了许多生词 (也可以说是新闻语料中有很多的生造词、简缩词), 另一方面是因为在语法词典的制作过程中, 这两条信息的限制很严, 对于真实语料中的定中式动名语串认识不够, 单凭人的记忆与经验自然不可能将所有的语言现象都列举出来。

### 3.2 应用语义词典中描述的动词与名词之间的语义选择限制信息

国内外对汉语语义分类体系的研究已有了一些成果, 如: 梅家驹《同义词词林》, 林杏光《简明汉语义类词典》, 陈群秀、张普《信息处理用现代汉语分类体系》, 王惠、詹卫东、刘群《〈现代汉语语义词典〉<sup>12</sup>的概要及设计》。本文将采用王惠等所设计的《现代汉语语义词典》的语义分类体系<sup>13</sup>。

《现代汉语语义词典》对常用实词的语义分类及语义搭配信息作了描述, 1998 年已经完成 4 万余词的语义描述。该词典“从工程实用的目的出发, 选择配价理论作为语义分析的理论框架”, 主要对“主体”、“客体”、“邻体”等三种配项成分进行语义限制。其中, 主体指动作行为的发出者或性状的承担者, 客体指动作行为所涉及的对象或性状的关涉对象。

本文试图引入该词典所描述的语义信息来判定“v+n”结构是否为定中关系。在此, 我们着重分析傅承德(1993)中提出的无法解决的实例。

根据邢福义(1994), 汉语动宾关系的复杂情况主要发生在单音动词和宾语之间, 双音动词和名词之间, 一般都是动词带直接对象宾语或间接对象宾语(目标宾语)。工具、处所、模式类宾语很少接在双音节动词后面。我们主要考察双音节动词直接修饰名词的现象, 所以, 在判定过程中, 我们假定述宾结构的宾语的语义类应当与述动词的客体语义类相一致, 否则即应当为定中结构。至于其中的交叉部分, 我们将另文考虑。

词语语义描写格式为:

V (主体语义类, [客体语义类|客体语义类])<sup>14</sup>,

N (语义类)。

如果前面已经对某词的语义信息进行过描述, 则后面的从略。

判定原则(分析成功或者分析失败均相对于原结构关系而言, 即推导出的结构关系与实际结构关系相符即为分析成功):

如果 V 的客体语义类与 N 的语义类相容(N 的语义类与 V 的客体语义类相同或者为其子类), 且自动判定为述宾结构, 则结果正确;

<sup>12</sup> 该词典在 1998 年已经完成对 4 万余条常用实词的语义分类及搭配信息的逐一描述。该文主张将语义分析作为一种辅助方法, 协助解决句法分析所无法解决的歧义问题。

<sup>13</sup> 参见王惠等(1998)。

<sup>14</sup> “[ ]”内的内容为可选内容, “[ | ]”相当于或者。

如果 V 的客体语义类与 N 的语义类不相容，且自动判定为定中结构，则结果正确；

如果 V 的客体语义类为空<sup>15</sup>，且自动判定为定中结构，则结果正确；  
其它，结果错误。

### 3.2.1 实例分析<sup>16</sup>

#### A 类动词分析

- (1) 印刷（人类，作品）工厂（集体）=>定中关系，正确
- (2) 流行（人为事物/抽象事物）服装（服饰）=>定中关系，正确
- (3) 教育（人类，人类）经费（人为事物）=>定中关系，正确
- (4) 建筑（人类，建筑物）材料（抽象事物）=>定中关系，正确
- (5) 装订（人类，作品）车间（具体空间/集体）=>定中关系，正确
- (6) 包装（人类，具体事物）机器（用具）=>定中或述宾关系，？
- (7) 印刷传单（作品）=>述宾关系，正确
- (8) 流行肺病（生理）疾病（生理）=>述宾关系，正确
- (9) 教育孩子（亲属和关系）=>述宾关系，正确
- (10) 建房房屋（具体空间/建筑物）=>述宾关系，正确
- (11) 装订书刊（作品）=>述宾关系，正确
- (12) 包装成品（人为事物/抽象事物）=>述宾关系，正确

#### B1 类动词分析：

- (13) 创作（人类，作品）小说（作品）=>述宾关系，正确
- (14) 投递（人类，作品）信件（作品）=>述宾关系，正确
- (15) 欢迎（人类，人类/人为事物/抽象事物）来宾（身份）=>述宾关系，正确
- (16) 广播(人类，信息)文章（作品）？=>
- (17) 汇报（人类，抽象事物）思想（意识）=>述宾关系，正确
- (18) 创作时间（时间）=>定中关系，正确
- (19) 投递业务（事情）=>定中关系，正确
- (20) 欢迎仪式（事情）=>定中关系，正确
- (21) 广播器材（用具）=>定中关系，正确
- (22) 汇报会议（抽象事物）=>定中关系，错误

#### B2 类动词分析：

- (23) 认识（人，人）问题（作品）=>定中关系，正确

<sup>15</sup>即该动词为不及物动词。

<sup>16</sup>分析结果分三种：正确、错误、存疑（用？表示）。

- (24) 认识过程(抽象事物) =>定中关系, 正确  
 (25) 认识水平(事理) =>定中关系, 正确  
 (26) 认识动物(动物)  
 (27) 认识世界(抽象空间)  
 (28) 认识大伙(?人<sup>17</sup>) =>述宾关系, 正确

#### C类动词分析:

- (29) 分析(人类, 抽象事物)质量(抽象事物) =>定中或述宾关系, ?  
 (30) 检查(人类, 人为事物/抽象事物)内容(作品?) =>定中或述宾关系, ?  
 (31) 评论(人类, 人类/抽象事物)方式(事理) =>定中或述宾关系, ?  
 (32) 整理(人类, 事物~时间~人类)方法(事理) =>定中或述宾关系, ?  
 (33) 培养(人类, 人类)目标(事理) =>定中关系, 正确  
 (34) 认识(人, 人)对象(抽象事物) =>定中或述宾关系, ?

#### 3.2.2 “Vn+n”语串分析

由于动词的语义搭配信息和名词的语义类主要是用于处理句子层面的主谓、述宾结构的, 所以我们可以假定这些信息用于测试述宾式动名短语时正确率应该相当高, 所以在此我们的注意力主要集中于定中式动名短语上。我们从标注语料库中随机抽取出 3023 例“Vn+n”(即定中结构)的语串, 而后用《现代汉语语义词典》中描述的动词客体语义类和名词语义类信息对 3023 例语串进行分析之后, 得出如表二所述的数据:

表二

动词部分未描述数	动词部分已描述数		
	分析成功数		分析失败数
336	动词为不及物数	动词为及物数	594
	961	1132	

如表二所示:

3023 个语串中有 336 个语串的动词部分在语义词典中无对应项, 即没有得到描述。分析成功的有 2093 项。在 2093 项中, 有 961 个语串动词客体语义类描述信息为空白, 即此类动词为不及物动词。分析失败的有 594 项, 所占比率为  $594/2687=22.1\%$ 。根据我们所用的判定原则, 这 594 项均为动词客体语义类与名词语义类信息相一致或后者为前者的子类。以后名为“部门”的几个语串为例:

<sup>17</sup> 此词尚未得到语义描述, 暂定为“人”。

表三

动词	名词	客体语义类	主体语义类	名词语义类
承办	部门	事物	人类	集体 人类 生物 具体事物 事物
管理	部门	具体事物	人类	集体 人类 生物 具体事物 事物
监察	部门	人类	人类	集体 人类 生物 具体事物 事物
监督	部门	人类	人类	集体 人类 生物 具体事物 事物
监管	部门	人类	人类	集体 人类 生物 具体事物 事物
教育	部门	人类	人类	集体 人类 生物 具体事物 事物
领导	部门	人类	人类	集体 人类 生物 具体事物 事物
运输	部门	具体事物	人类	集体 人类 生物 具体事物 事物
组成	部门	具体事物	具体事物	集体 人类 生物 具体事物 事物

如上所示，“部门”原语义类为“集体”，是“生物”、“人类”、“具体事物”、“事物”的子类，从而与上表中的动词的客体语义类相一致。总体上说，语义词典中对于动词的主客体的语义类描写较粗一些，所以限制也较松，从而导致了上述的错误判断。

我们认为本文所用的语义词典的分类体系颇有不足，尤其是在用于解决本文中的问题时表现得更加明显。像“部门”这样的词，属于较虚的词，即便它有集体的意义，也是概括某一类事物的名词，且可以分成许多子类。这一类名词前有动词时，倾向于直接受动词修饰。又比如“成本”被描述为“人为事物”，“单位”被描述为“集体”，与“部门”类似，都是不妥的。

### 3.2.3 后名动词与前动名词组成的述宾结构动名语串分析

对“句法属性”一节中由具有后名属性的动词与具有前动属性的名词组成非定中结构动名语串的479个实例，使用与上面相同的方法分析之后，得出的结果为：

分析成功的有258项，即召回率为53.9%；另222项中，80项为语料库标注错误，即应属于定中结构<sup>18</sup>；属于述宾结构的为52项，其它90项为跨结构的语串。

如果我们认为上面分析成功的258项和标注错误的80项（通过本节中的分析已得到成功分析）均属于判断成功的实例，那么，我们对于述宾结构的动名语串判断的成功率即为 $(7339-142)/7339=98.1\%$ 。

### 3.3 实验小结

前面，我们分别对从语料中抽取的“v+n”和“Vn+n”语串进行了分析实验。最后得出的结果为：以动词的后名属性与名词的前动属性相结合判定“v+n”语串为述宾结构的成功率为93.5%，再结合语义词典的描述信息判定则成功率提高到98.1%。但由于对定中式动名短语的判定正确率不足60%，结合起来的结果自然很不理想。单纯以语义词典的描述信息来判定“Vn+n”语串为定中结构的成功率为78.9%，这一结果也让人无法乐观。

<sup>18</sup> 由于时间和技术上的原因，这些错误的实例没有回原文核对，仅凭个人的判断。

究其原因,主要在于以下两个方面:(1)语法词典中对于动词的“后名”属性和名词的“前动”属性的描述很不可靠,这自然是因为缺乏大语料库支持的原因。但即便有了大语料库的支持,这一结果也不可能有太大改观。显然,具有这两种属性的动词和名词增加,对定中式判定正确率的提高必然伴随着对述宾式判定成功率的降低。因此,句法信息不足以识别这两种结构关系。(2)语义词典的名词语义类信息和动词配价信息是针对句子一级的,因而主要适用于述宾、主谓两类结构,对定中式动名短语并不太适用。

### 3.4 判别过程中存在的问题及下一步的打算

应当说,我们在展开研究的过程中,发现了一些问题,但是,由于时间的不足和个人能力的局限,很多问题没有能够进一步深入。接下来,我们打算就以下几个方面展开进一步讨论:

(1)完善语法词典中动词的后名属性和名词的前动属性,并在此基础上继续进行实验。

(2)以名词为中心,对定中结构动名语串的语义模式进行分类,可考虑语义场问题。这个可以与(1)结合起来进行。

### 参考文献

- 董振东、董强,2000,《关于知网——中文信息结构库》, [www.keenage.com](http://www.keenage.com)。
- 傅承德,1993,《论现代汉语动名语串的句法结构和语义关系的自动识别》,《语言研究》1993年第一期,32—44页。
- 顾阳、沈阳,2001,《汉语合成复合词的构造过程》,《中国语文》2001年第2期。
- 李晋霞,2002,《面向计算机的“V双+N双”结构类型研究》,《语文文字应用》第4期,69-76页。
- 王惠等,1998,《〈现代汉语语义词典〉概要及设计》,《1998中文信息处理国际会议论文集》,清华大学出版社。
- 邢福义,1994,《NVN造名结构及其NV|VN简省形式》,《语言研究》1994年第2期。
- 杨成凯,1992,《广义谓词性宾语的类型研究》,《中国语文》1992年第1期。
- 姚振武,1996,《汉语谓词性成分名词化》,《中国语文》1996年第1期。
- 俞士汶(主编),1999,《现代汉语语料库加工——词语切分与词性标注规范与手册》,北京大学计算语言学研究所。
- 俞士汶等,1998,《现代汉语语法信息词典详解》,清华大学出版社。
- 詹卫东,2000,《面向中文信息处理的现代汉语短语结构规则研究》,清华大学出版社、广西科学技术出版社。
- 詹卫东等,1998,《基于词组本位语法的语义模型》,《中文与东方语言信息处理学会学报》1998年第1期。
- 张伯江,1994,《语类活用的功能解释》,《中国语文》1994年第5期。

## Constitutive Relation Analysis for V+N Phrases

Qiu Likun

Institute of Artificial Intelligence, Beijing City University, Beijing, 100083

### Abstract:

An important and significant subject in Chinese POS tagging and syntactic analysis is to choose appropriate tags for verbs between finite and nonfinite forms. As far as V+N phrases are concerned, through choosing tags for verbs, at the same time, we may judge the constitutive relation of the phrases. This paper uses electronic syntactic dictionary and semantic dictionary to solve this problem and has got inspiring results.

Through twenty years' effort, electronic language resources have become more and more abundant. Now there are several kinds of electronic linguistic dictionaries including syntactic dictionaries and semantic dictionaries. Some of them have arrived at enough scale. In this paper, a syntactic dictionary and a semantic dictionary, both provided by the institute of computational linguistics, Peking University, are used.

First, we use two syntactic attributes of the syntactic dictionary, that is, the “后名” attribute of verbs and the “前动” attribute of nouns. We use these two attributes to judge the constitutive relation of V+N phrases. From the annotated corpus, we get 5303 instances in which verbs are nonfinite. Looking up the syntactic dictionary, we can see that their attributes are as follows:

the “后名” attribute of verbs	True	True	False	False
the “前动” attribute of nouns	True	False	True	False
Quantity	479	1105	935	2784

That is to say, we can get a precision of 93.5%. But when the verb of V+N phrases is finite verb, we only can get a poor precision of about 60%.

Second, the information of verb's dative valence and the classification information of nouns are utilized. Here, we get a precision of 78.9%.

**Keywords:** V+N phrases, constitutive relation. automatic analysis, contemporary Chinese, annotated corpus, electronic dictionary