汉语复合名词短语特征结构的标注研究*

万 菁[1], 姬东鸿[1,2], 任 函[1,2], 冯文贺[1]
[1]武汉大学 语言与信息研究中心, 武汉 430072
[2]武汉大学 计算机学院, 武汉 430072
E-mail: jennifer_wanj@yahoo.com; donghong_ji2000@yahoo.com.cn; hanren@whu.edu.cn; wenhefeng@gmail.com

摘 要: 复合名词短语的特征结构标注是基于特征结构表示的汉语大规模语义资源建设的一个子任务。本文探讨了在标注的过程中建立的最小关联原则、直接关联原则、语言关联原则等主要原则，同时也涉及复杂名词短语内部成分之间的语义关联种类的确定，这将有助于探讨适合汉语实际的语义表示机制及有效的汉语语义分析策略。

关键词: 特征结构; 名词短语; 概念关联; 关联种类

Study on Feature Structure Tagging for Chinese Nominal Compounds

Wan Jing[1], Ji Donghong[1,2], Ren Han[1,2], Feng Wenhe[1]
[1] Center for Study of Language & Information, Wuhan University, Wuhan 430072
[2] Computer School, Wuhan University, Wuhan 430072
E-mail: jennifer_wanj@yahoo.com.cn; donghongji@whu.edu.cn; hanren@whu.edu.cn; wenhefeng@gmail.com

Abstract: Tagging Chinese nominal compounds with feature structure is a sub-part of the semantic resource based on feature structure. The principles of Minimalist relatedness, directly relatedness, language relatedness, efficiently relatedness and consistently relatedness, are primarily about what to label, as well as the problem of how and why to tag.

Keywords: feature structure; nominal compounds; concept relatedness; relatedness type

1 前言

语义分析是现代语言学和计算语言学领域最具挑战性的研究之一，也是当前制约语言信息技术大规模应用的主要瓶颈。语义分析的首要任务是确定要获取什么样的语义信息。本文引入特征结构的概念，旨在分析汉语复合名词短内部结构及其语义关系，探讨有效的汉语语义分析策略。

特征结构在现代语言学并不是一个新术语。语音学很早就采用类似特征结构的机制描述音节，后来形式句法理论如 GPSG 和 LFG 又采用复杂特征集描述句法结构，复杂特征集也类似特征结构。这两种情况都是定义一组特征用以区分音节或句法结构，分别在生成语音学和生成语法领域产生了很大影响。可是至今为止，还未见到利用特征结构进行大规模的语义描述及语义分析的尝试。

2 特征结构及复合名词短语

2.1 特征结构

在以往的研究中，我们发现，有些短语我们虽然知道其内部成分之间存在关联，但是并不清

* 本文获国家自然科学基金重大研究计划“视听觉信息的认知计算”培育项目“汉语特征结构的资源建设和自动分析研究”（90820005）、国家自然科学基金面上项目“基于部分指导的词义学习和词义排歧综合研究”（61070082）、2008-2009 年度武汉大学人文社科自主创新项目“基于语义的网络舆情智能监测平台研究”、武汉大学 985 二期拓展子项目“基于汉语特征结构的语义描写及其应用”（985yk004）、湖北省教育厅人文社科项目“基于依存语法的语料库标注研究”（2008q275）资助。

楚哪些成分之间存在关联，或各成分之间存在怎样的关联（关联种类）。参照 HowNet 等词汇语义知识库中对词语进行的分析，可以确定，在“黑色皮鞋”这个短语中，“皮鞋”作为一个实体，“色”是“皮鞋”的特征，“黑”是“色” 的特征值，那么我们可以用一个三元组的形式表示他们的关系，即【皮鞋，色，黑】，从某种意义上说，特征就是连接实体与特征值的关系。由此，我们可以认为，一个短语的三元组集合就为该短语的特征结构。

形式上，一个三元组可看作两个点和连接它们的边，其中的节点表示实体和特征值，边表示特征。于是一个特征结构可看作一个依存图。考虑到特征值也可是另外一个特征结构，因此特征结构可看作一个递归依存图，即节点本身又可是一个图。依存中心为短语成分关系命名的参照点，通常是关系对中的语义重点。

2.2 复合名词短语

复合名词短语通常是由一串名、动、形等实义词构成的名词短语，如：工程施工招投标管理办法、省级经济管理权限。过去对于名词短语的分析主要集中在 NN 这类名词短语的语义解释上，主要任务是自动获取修饰语和中心语之间隐含的语义关系。通常是采用两种处理策略，一种是首先定义一组关联类别的集合，然后为每个名词短语分配适当的关联集合，即自上而下的策略。另一种是不定义短语内部成分之间的关系而通过大规模语料去发现词语组合时隐含的语义关系，即自下而上的测略。而汉语语言学的分析更是限制在一些特定的名词短语格式上，如 NVN、NV、VN、NN 等。

另一方面，即使在一些大规模的真实语料库标注中，由于不同于典型句法结构，复合名词短语往往也未能达到恰当处理。举以层级关系为基本结构原则的宾州 CTB[7]来说，由于难于确定词语间的关系，众多复合名词短语通常处理为平头结构，如 ((工程)(施工)(招投标)(管理)(办法))、((省级)(经济)(管理)(权限))。这样的结构分析，显然不能满足语义关系处理的需要。

考虑到以上几点，我们标注的复合名词短语均来自宾州 CTB，数据规模大约为 20000 个复合名词短语。并在确定复合名词短语的范围时我们考虑了下面的一些因素：

第一，不含“的”的复合名词短语。这一条主要是为了排除定语为小句的情况，因为其结构方式和一般复合名词短语的结构规律差异很大。如：国家开发银行今年发行的六百五十亿金融债券。这种情况下，去掉“的”名词短语就成了一个普通的句子。而对于如：重庆的国民经济和社会事业、中国人民银行的贷款。我们认为去掉其中的“的”字后，其结构方式和普通名词短语的结构方式并无大的差异，虽暂时未列入标注范围，但拟在后续的名词短语标注中进行标注。

第二，没有区分专有名词短语与非专有名词短语。看这两个名词短语：陕西省外商投资企业管理办法、《陕西省外商投资企业管理办法》。仅因为有无书名号的差异，也许一个要作为专名处理，一个要作为普通名词短语处理。

其实如果仅考虑名词短语内部词语间的关系，上面的这些关系可以统一作为同样的名词短语处理。是否“专名”其实是一个“使用”问题，与名词短语内部结构无任何关系。另一方面，假如一个“专名”并非普通名词短语，那么它不在我们的分析范围内，如：《中国向何处去？》。

第三，包含三个或三个以上构成成分。考虑到仅包含两个词的复合名词短语内部结构关系比较单一，我们的标注对象包含三个或三个以上构成成分。

3 概念关联及关联原则

3.1 直接关联原则和最小关联原则

在确定了我们标注的范围后，我们亟待关注的就是复合名词短语各个概念之间的关联原则。

概念关联就是短语或句子中词汇概念[1]之间的联系。以1)为例：

1) 工商银行贷款利息

我们认为，“工商”和“银行”、“银行”和“贷款”以及“贷款”和“利息”间存在着概念关联，而且这种关联是直接的、最小的。

那么我们在标注中就产生了两类标注原则，即**“直接关联原则”**和**“最小关联原则”**。因为，通常情况下，“银行”和“贷款”，“贷款”和“利息”之间是存在关联的，而“银行”和“利息”之间有时也被认为存在关联，但不同的是，“银行”和“利息”之间是一种间接的关联，可以通过“银行”和“贷款”，“贷款”和“利息”推出来。所以我们就只标注具有直接关联的两个概念。其原因有二：

第一，通过直接关联可以定义或推理得出间接关联乃至所有关联。从科学性来说，这可以更准确的揭示概念之间的关联。

其实就哲学意义上来说，世界万物之间都有关联。但对于人的认识尤其是科学研究来说，揭示直接关联（或者说是基本关联）才能更深刻、准确的理解、认识万物之间的关联。比如，通过“婚姻关系”、“生育关系”等几个基本关系人就可以定义或理解所有的错综复杂的亲属关系 。由直接关联而认识间接关联，既是一种科学研究的方式，也反映了人认知世界的普遍方式。

第二，从可行性来说，仅标注直接关联可以减少标注的关联数，是一种经济可行的做法。

另外，此例中是“银行”与“贷款”，“银行”与“利息”发生关联，而不是“工商银行”整体与“贷款”和“利息”发生关联，这就是所谓的最小关联。原因在于：

第一，最小关联具有确定性，而整体关联通常具有多种切分的可能性。又如“上海浦东开发银行”中，“上海浦东开发”具有两种整体关联的可能1)【上海浦东，，开发】【浦东，，上海】；2)【浦东开发，，上海】【开发，，浦东】。根据不同的语境和不同人的语感可能会选择不同的切分与关联。

第二，最小关联实际可以推出整体关联[2] 。对比下图：

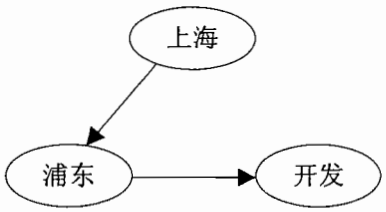


图1(a) 上海—浦东—开发

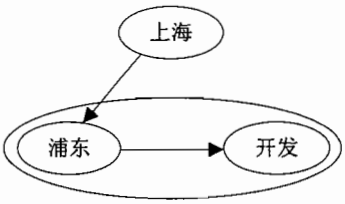


图1(b) 上海-浦东—开发

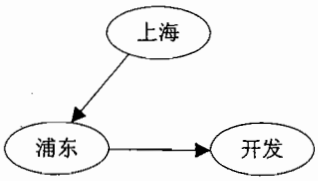


图1(c) 上海—浦东-开发

[1] 下文可看出，也可能是复合概念。

[2] 短语结构语法分析实践告诉我们，整体结构（短语）是大小长短不一的单位，理论上从词到句子之间的任何一个组合都可以称为短语，短语结构的关联就发生在这些因语境、因个人语感不同而切分出来的大小不一的整体结构之间。

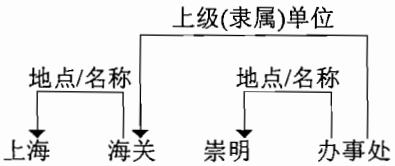
图 1(a)是最小关联，“浦东”与“上海”“开发”均有关联。图 1(b)显示，“开发”可以与“上海浦东”整体发生关联是因为“开发”可以与“上海浦东”的 “浦东”发生关联；同理，图 1(c)显示，“上海”可以与“浦东开发”整体发生关联是因为“上海”可以与“浦东”发生关联。

第三，假如我们一开始就进行整体关联分析，然后通过对整体进行逐层分解得到较小关联或最小关联，那么就可能得不到准确的最小关联或失去一些最小关联[1]。与最小关联比，整体关联不仅具有不确定性，而且还具有非准确性。

最小关联的确定性决定我们的标注可以获得较高的标注一致性；最小关联对整体关联的推测性决定了我们仅通过最小关联就可以获得整体关联。这是我们确定最小关联标注原则的根本原因。

3.2 语言知识关联原则

首先来看下面的例 2）上海海关崇明办事处，

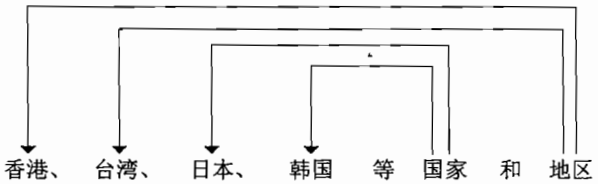


直观上，人们会对 2）的语义有如下理解：

“办事处”的【地点】及【名称】均是【崇明】，“办事处”的【上级/隶属单位】是“海关”，“海关”的【地点】及【名称】均是“上海”。这里我们仅认定【办事处， ，崇明】这一关联，而不认定【崇明， ，上海】这一关联（尽管现实生活中“崇明”确实是“上海”的一个辖区）。这是因为后者之间的关联需要一定的地理知识才能判断，而前者仅需正常人的语言知识就能做出判断。

再来看下面的例子：

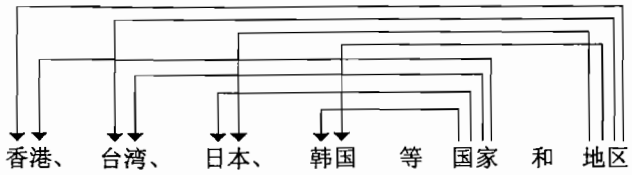
3）香港、台湾、日本、韩国等国家和地区



这里我们无法认定“香港”“台湾”“日本”“韩国”等分别和“国家”、“地区”之间的关系分配。可以设想这样的情况，比如让一个熟练掌握了语言的儿童来判断，又比如让一个完全不懂政治的人来判断，或比如让处于不同历史时代的人来判断，得到的其间的关联也许完全不同。这样的关联分配显然也已经超出了一般人的语言能力的判定范围。不过，在这种情况下，可以肯定的是正常人的语感能判定“香港”“台湾”等一定和其后的“国家”或“地区”发生某种关联。鉴于此，我们采用了下面的关联判定：

---

[1] 通常认为短语结构可以等价地转换为依存结构，这种看法建立在“中心”可以作为结构整体的代表与外界发生联系的假设上[5]。不过这种假设有时候并不成立，考虑上面图 2B2'，经过这种转换以后“上海”会与“浦东开发”的中心“开发”发生关联。显然，“上海”与“开发”的关联并不准确。又如“大规模杀伤性武器”，假如分析为：（（大（规模））（杀伤性（武器））），经过转换以后“大规模”的中心“规模”将于“杀伤性武器”的中心“武器”发生关联。这种关联显然也是不准确的。有关依存结构与短语结构的语义关联的等价性问题我们将另文详细探讨。另外需要指出的是“中心”在特征结构理论中并不是一个必须的基本概念[1]。



4 概念关联及关联种类

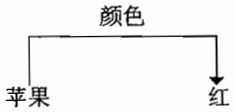
在确定了复杂名词短语内部的关联后，就需确定这些概念间存在的关联种类。关于概念种类的理解，我们也通过一例来解释。如下例中

4）从北京飞到上海

“飞”与“北京”、“上海”都有一定关联，可以表示为【飞，北京】【飞，上海】这样的概念关联对。如果把关联的种类加到概念关联对中，上面的概念关联对可表示为：【飞，起点，北京】【飞，终点，上海】。我们可直接用“从”和“到”这两个虚词表示它们之间关联的种类。但有时，关联的种类并不一定出现，如“红苹果”和“红色苹果”比较，它们的特征结构分别为【苹果，，红】和【苹果，色，红】，我们就需要在其图结构中进行补足，他们的特征结构依存图可都表示如下：

5）a. 红苹果

b. 红色苹果

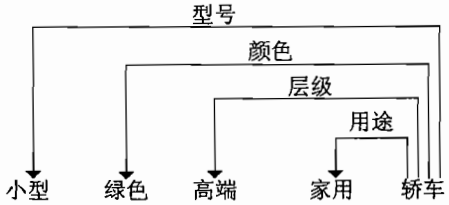


对关联种类的确定需要解决如下几个问题：

第一，关联种类的概括。

这里需要研究的是如何建语义关联类。看下例：

6）小型绿色高端家用轿车



由上例可以看出，语言中抽象的关系词往往可以作为关系类的代表，只不过有些情况下可能会嵌入到特征值中，如“型号”嵌入到了“小型”中，“颜色”嵌入到了“绿色”中。

这里的关键问题是如何从一般意义下的特征词中分离出特征，并给特征分类，以作为关联种类。

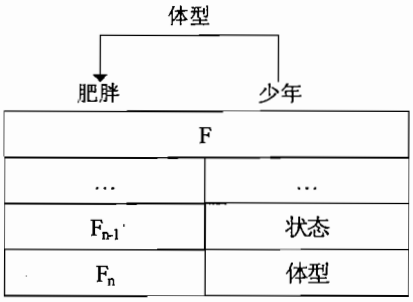
第二，关联种类的选择。

概念上，关系类是唯一的，但类名却可以很多，并且实际上同一个类在现实生活中往往可用不同的词汇来表达。比如“型号，型，号，号码，码”“大小，尺寸，尺度”等等，表达的关系类可能是相同的。另外，还有一些关系，可能并不能用一个词来概括，比如说“妇女小说”中“妇女”可能是“小说”的【描写对象】或【内容】等。我们在这里的处理策略是在关联类概括的基础上进一步选择出有代表性好记易用的关联种类名，同时给出同类的词的集合（同类词群）。

第三，建立具有层级的关系类结构。有一些研究值得参考，如知网的关系系统。对于这一部

分，我们采取的方法是将所有关系类分成若干层级，最终形成一个层级结构，每个类中，下层关系遗传了上层关系的所有特征。表 1 是我们根据相关研究初步建立的关系类示例。

表 1　语义关系类层级结构图



总之，将会结合建设和建立和完善一个科学、合理、直观、易用的语义关系类及 Ontology。

5　结语

在标注过程中未能很好解决的问题有两个：

第一，最小关联单位的确定问题。要给出一个严格的“词”的本质或操作性定义是困难的。其实就问题的本质来说最小关联单位（概念词）的确定问题，与特征结构标注本身无关，“词”的判定分歧在其他的语法或语义理论中同样存在。

第二，关联种类的判定问题。对于某种概念关联其关联的种类可以是多个时，我们往往难以判定选用哪个种类更好。

参 考 文 献

[1] N. Xue, F. Xia, The Bracketing Guidelines for the Penn Chinese Treebank (3.0), 2000.
[2] 姬东鸿. 汉语特征结构的资源建设和自动分析研究. 国家自然科学基金申请报告[R], 2008.
[3] 赵军，黄昌宁. 基于转换的汉语基本名词短语识别模型[J]. 中文信息学报，1999(2).
[4] 邢福义. NVN 造名结构及其 NV丨VN 简省形式[J]. 语言研究，1994(2).
[5] 车竟. 试论“N十V”式定心结构[J]. 汉语学习，1994(1).
[6] 李晋霞. 现代汉语动词直接做定语研究[M]. 北京：商务印书馆，2008 .
[7] 谭景春. 名名偏正结构的语义关系及其在词典释义中的作用[J]. 中国语文，2010(4).
[8] 朱德熙. 语法答问[M]. 北京：商务印书馆，1985.
[9] 朱德熙. 句法结构(1962). 载现代汉语语法研究[C]. 北京：商务印书馆，1980.
[10] 冯志伟. 机器翻译研究[M]. 中国对外翻译出版公司，2004.