文章编号：1003-0077（2018）00-0000-00

中文基本复合名词短语语义关系体系及知识库构建[1]

刘鹏远，刘玉洁

（北京语言大学 信息科学学院，北京 100083）

摘 要：名词短语一直是中外语言学领域的重要研究对象，近年来在自然语言处理领域也受到了研究者的持续关注。英文方面，已建立了一定规模的名词短语语义关系知识库。但迄今为止，尚未建立相应或更大规模的描述名词短语语义关系的中文资源。本文借鉴国内外诸多学者对名词短语语义分类的研究成果，对大规模真实语料中的基本复合名词实例进行试标注与分析，建立了中文基本复合名词短语语义关系体系及相应句法语义知识库，该库能够为中文基本复合名词短语句法语义的研究提供基础数据资源。目前该库共含有 18218 条高频基本复合名词短语，每条短语均标注了语义关系、短语结构及是否指称实体等信息，每条短语包含的两个名词还分别标注了语义类信息。语义类信息基于北京大学《现代汉语语义词典》。基于该知识库，本文还做了基本复合名词短语句法语义的初步统计与分析。

关键词：基本复合名词短语；语义关系体系；知识库

中图分类号：TP391　　**文献标识码：**A

Semantic Relations Hierarchy and Knowledge Base Construction of Chinese Basic Noun Compounds

LIU Pengyuan, LIU Yujie

（Beijing Language and Culture University,Beijing 100083）

Abstract: Noun compound has always been important in the field of linguistics both in China and abroad. In recent years, researchers of natural language processing have been paying close attention to the noun compound too. In English, a relatively large-scale noun compound semantic relation knowledge base has been established. But so far, the corresponding Chinese resources describing the semantic relations of noun compounds have not been established. Based on the research of the semantic classification of noun compounds by many researchers, this paper tries to tag and analyze the basic compound nouns in the large-scale real corpus, and establishes the basic noun compound semantic relation hierarchy and the corresponding syntax and semantic knowledge base in Chinese. We hope that this knowledge base can be used as a fundamental data resource for the research of basic noun compound. At present, the knowledge base contains 18218 high-frequency basic noun compounds, each of them is labeled with semantic relation, phrase structure and referential entity information. Each noun compound contains two nouns with semantic category respectively. The semantic category is based on the SKCC of Peking University. Based on this knowledge base, we also made preliminary statistics and analysis of syntactic and semantics of basic noun compounds.

Key words: noun compound; semantic relations hierarchy; knowledge base

---

[1]**收稿日期：**　　**定稿日期：**

基金项目：教育部人文社科规划项目（18YJA740030）

作者简介：刘鹏远（1974.5），男，博士，主要研究方向为自然语言处理；刘玉洁（1994.1），女，硕士生，主要研究方向为计算语言学。

1 引言

名词是人类语言最基本的词类范畴之一，包含了大量思维和认知信息，在语言学研究中占有非常重要的地位。对名词短语，不但一直是中外语言学领域的重要研究对象，近年来在自然语言处理领域也受到研究者的持续关注。

国内外学者很早就对名词短语语义关系分类进行研究。国外，Downing[1]针对英语复合名词短语提出了十二类语义关系。国内，吕叔湘先生[2]将作定语的修饰成分名词与中心名词之间的关系分成三大类。随后，众多学者针对名词短语提出了很多语义关系分类方法，但鲜有基于语料库大规模实例的验证。

Vanderwende[3,4]首先进行了名词短语语义关系标注知识库建设的尝试，但规模很小。目前最大的英语复合名词短语语义关系知识库是 Tratz&Hovy[5]建立的，该库共含有 17509 条短语，标注了十二类语义关系，每一个大类关系下，还分了小类，也进行了语义关系标注。迄今为止，汉语并无类似的开放语义资源，仅有魏雪和袁毓林[6,7]以隐含谓词的识别和自动释义为目的而建立的名名搭配知识库，规模为 638 条，暂没有开放。

建立复合名词短语语义关系体系，建立该体系下具有一定规模的知识库，可以帮助研究者们分析及发现名词短语的句法语义规律，对名词短语的自动释义、语义关系自动分类及名词短语复述等相关任务的研究具有很大价值。本文针对基本复合名词短语[2]，对中外众多学者提出的语义分类体系进行梳理，并以 Levi[8,9]的语义分类体系为基础，结合汉语相关研究及汉语自身特点，在反复考察大规模语料库中基本复合名词短语实例的基础上，建立了中文基本复合名词短语语义分类体系；利用该体系，对 18218 条新闻领域高频基本复合名词短语进行了人工标注。标注内容为：语义关系，句法结构，语义类及是否实体指称等信息。

本文后续组织如下：第 2 节对相关研究进行了综述；第 3 节介绍 NN 短语的语义关系分类体系以及知识库构建；第 4 节是对建立的知识库的语义基本情况的考察；最后一节对全文进行总结。

2 相关研究

2．1 复合名词短语语义关系相关研究现状

国外复合名词短语语义关系分类的研究主要有两条路线，其一是通过复合短语内部各个成分的语义类来定义其语义关系(始于 Downing[1])，另一条则是通过对删除谓词的语义类来定义复合名词短语内部成分的语义关系(始于 Levi[8]及 Warren[10])。

Downing[1]认为对于某一个有限数量的语义或语法种类的关系可以作为短语关系的潜在关系。他提出了十二类语义关系：(1)Whole-Part；(2)Half-Half；(3)Part-Whole；(4)Composition；(5)Comparison；(6)Time；(7)Place；(8)Source；(9)Product；(10)User；(11)Purpose；(12)Occupation。

Levi[8,9]在对英语复合名词短语的研究中，通过删除谓词获得的“ N1+N2”复合名词短语的名词成分之间的语义关系进行分类，并根据可删除谓词的语义类，提出了此种结构中作修饰成分的名词和核心名词之间的十二种语义关系：(1)N1 CAUSE N2；(2)N2 CAUSE N1；(3)N1 HAVE N2；(4)N2 HAVE N1；(5)N1 MAKE N2；(6)N2 MAKE N1；(7)N2 USE N1；(8)N2 BE N1；(9)N2 IN N1；(10)N2 FOR N1；(11)N2 FROM N1；(12)N2 ABOUT N1。

Warren[10]认为复合名词短语的一个特点就是其中的抽象语义关系，这种语义关系由四个层级组成，最顶层由六种粗粒度的语义关系，分别为(1)Possession；(2)Location；(3)Purpose；(4)Activity-Actor；(5)Resemblance；(6)Constitute；而每一大类下面又包含众多细粒度语义关系类型。

随后的研究者在以上研究基础上继续改造或细化。陆续有[11-14,5]等。

国内传统汉语相关研究中，复合名词短语内部名词之间的关系集中在修饰语（定语）和中心语的关系上，相关研究常常面向包括了“N1+（的）+N2”结构以及其他名词中间插入了其他成分的结构，统一将之看作名词作修饰语（定语）的情况进行讨论。吕叔湘[2]将作定语的修饰成分名词和中心名词的关系分为领属性的，描写性的，同位性的；朱德熙[38]

[2] 复合名词短语源于英语研究者所用“Noun Compounds”，意为多个名词复合而成的短语，整体表现类似一个名词。本文将该结构的名词个数限定为 2，这样的结构是复合名词短语中最“基本”的一类，因此加上“基本”，对复合名词短语做词数限制。以下多将简称“NN 短语”。

提出修饰语和中心语意义上的联系是多种多样的，主要包括：表示领属者，表示质料，表示时间，表示处所等等。

相关研究成果如按照分类数量分，可分为两类。一类是两类说(两大类，大类下可能再分为小类)：袁毓林[15]将名词作定语的情况分为领属定语和属性定语两大类；张卫国[16]将名词作定语的情况分为限定性和区别性两大类；李宇明[17]将“N1 的 N2”的结构中，名词之间的关系分为属性关系和非属性关系；文贞惠[18]将“N1（的）N2”结构中名词之间的语义关系分为领有范畴和属性范畴两大类；后续还有蔺璜[19]及谭景春[20]的相关研究。另一类是多类说：黄国营[21]认为“N1 的 N2”中名词之间的语义关系共有十种：领属、属性、材料、比喻、同一、相关、成数、施事、受事、举例；孔令达[22]又进一步将其细分为十四类；单强、牛守祯[23]名将名词作定语的情况分为领属、数量、时间及处所定语等；马洪海[24]考察了“名+名”偏正结构和复指结构，将偏正结构语义关系分为七类,把复指关系的名名组合语义关系分为了八类；周日安[25]归纳出 18 种复合名词短语语义格组合；魏雪[26]归纳出了 26 种语义组合关系。

从语言信息处理角度，对应的是槽关系的研究，如鲁川[27]、林杏光和张庆旭[28]等。但更多的研究并不是聚焦在复合名词短语上，而是在汉语所有结构的语义关系层面进行的研究，如冯志伟[29]根据依存句法，提出了 30 种论元关系；鲁川[27]提出的意合网络种归纳出了 6 大类，共 26 种关系；董振东[30]提出事件内部语义关系总计 83 类，分为语义角色及辅语义角色；刘开瑛[31]基于 CFN 概括了 31 个常用的周边语义角色。

总的来说，前人的相关研究中，国外有成系统的短语内部名词之间的语义关系体系，在此基础上的研究成果颇丰。而国内的研究往往是针对名词作定语的情形中对名词定语进行的分类,而缺少对 NN 短语的针对性研究。在 NN 短语中，其结构类型也是多样的，不仅有定中结构，还有同位结构、主谓结构和联合结构。而针对 NN 短语的研究又缺乏相应的语义关系分类体系，仅仅以语义的组合来代替短语内部名词之间的语义关系，无法体现语义组合内部的深层关系。

2．2 名词短语语义关系相关知识库建设现状

国外最早由 B. Rosario&Hearst[32]建立了包含 1660 条名词短语及其语义关系的知识库。随后 Kim&Baldwin[33] 、 D Ó Séaghdha&Copestake[14] 及 Girju[13]分别构建了包含短语及其语义关系的知识库，规模分别为 2169、1443 及 2031 条。目前国外最大的名词短语语义关系知识库规模为 17509 条，是 Tratz&Hovy[5]构建的，标注了语义关系及名词词性。

严格说来，汉语目前尚无类似的知识库资源，特别是开放资源。仅有肖国政[34]建立基于语义依存图的汉语复杂名词短语资源，但没有规模等信息，没有开放。魏雪和袁毓林[6,7]以隐含谓词的识别和自动释义为目的建立的名名搭配知识库，规模为 638 条，没有开放。卢涌[35]针对“名词+的+名词”的结构总结出了三十多个释义模板，形成了一个短语释义库，规模为 1000 条，没有开放。其余语义关系知识库资源并非针对复合名词短语，词汇级别的主要有知网等，句子级别的有哈工大和北语联合开发的语义依存关系标注语料库等。

3 NN 短语语义关系体系及知识库构建

3.1 NN 短语获取

为保证 NN 短语的规范性，本文基础数据来源为新闻语料，选自国家语言资源动态流通语料库(DCC)2005 年至 2015 年的全年报刊数据，共超过 30 亿字次。我们使用 LTP 平台和 jieba 分词对语料进行分词及词性标注，以连续词性为“名词名词”的序列为识别模式，抽取上述两种分词结果得到模式的交集，最终得到 290 多万条 NN 短语，其中出现频次在 100 以上的有 22474 条。

在这些 NN 短语中，有小部分不属于本文考察范围，主要有三种情况：

(1)**多层嵌套名词结构的一部分**。如：“*《今日美国》报道，中国国家统计局公布的中国经济成绩单提振了全球市场信心*”。“中国国家”是形如“中国国家 XXX”的一部分，随着该类型结构的大量出现，导致“中国国家”频次较高。

(2)**分词粒度不一致**。如：“*其中，进城务工人员子女的教育，特别是义务教育问题日益引起了社会的关注*”。分词工具将“务工人员”视为一个名词，而实际上“务工人员”是两个词构成的短语，根据本文的定义，“务工人员子女”就不是本文关注的 NN 短语。

(3) **词性标注错误**。如：“*民警随即将犯罪嫌疑人王某龙及其车辆带回调查*”。分词工具将“犯罪”标注

成名词，而该词只有动词词性，故而“犯罪嫌疑人”非 NN 短语。

人工对 22474 条短语进行分析筛选和确认，共获得 18218 条 NN 短语。

3.2 语义关系体系

汉语相关研究中，并没有针对 NN 短语建立的语义关系体系。本文参考研究与应用最为广泛的 Levi 体系(英语)，然后结合汉语中对名词短语的相关研究成果，对上述 18218 条短语进行了反复试标注，由于汉语与英语 NN 短语存在很大差异，本文最终所建立的汉语 NN 短语语义关系体系与 Levi 体系有较大的差别。本文建立的语义关系体系共包含十四类语义关系：

（1） Cause：致使/引起/导致(因果关系)

N1 是导致 N2 产生的直接原因或 N2 是导致 N1 形成的直接原因。

①原因+结果：地震灾区（由于地震形成的灾区）、病毒感冒（由病毒引起的感冒）

②结果+原因：禽流感病毒（引起禽流感的病毒）、事故车辆（造成事故的车辆）

（2） Have：患有/含有/拥有(因果关系)

可以用“有”相关的词来解释。 N1 是 N2 的外部特征。

①患有：自闭症儿童（患有自闭症的儿童）、重症病例（患有重症的病例）

②拥有：技术人才（拥有技术的人才）、实体书店（有实体的书店）、武装分子（拥有武装的人）

③含有 4：碳酸饮料（含有碳酸的饮料）

（3） Make：用...做成/由...组成(做成组成关系)

N1 是 N2 的主要和直接组成成分。

① 材料成分：木头桌子（用木头做的桌子）、水果沙拉（主要用水果做成的沙拉）

② 成员：志愿者队伍（由志愿者组成的队伍）

（4） Use：使用...做/采用(使用关系)

N1 是 N2 的使用工具、使用方法和使用材料。可以通过“用”类相关动词来连接释义。

① 用工具：砂锅排骨（用砂锅炖的排骨）、钢琴协奏曲（用钢琴弹奏的协奏曲）

② 用原料：燃气热水器(使用燃气的热水器)、汽油发动机（使用汽油的发动机）、激光武器（使用激光的武器）

③ 用方式：法治中国（采用法治方式治理的中国）、冷链物流（使用制冷技术的物流）

（5）Be：是(属性说明关系)

N1 对 N2 某种属性的说明和补充，N1 和 N2 是从不同侧面对同一事物的描述。

① 指称：总统普京（总统[姓名]是普京）、英雄黄继光(英雄[姓名]是黄继光)、埃博拉病毒([名字]是埃博拉的病毒）、深圳特区（[名字]是深圳的特区）、东风汽车（[品牌]是东风的汽车）

② 补充：老王夫妇、夫妻双方、母女关系

③ 陈述：今天星期一（今天是星期一）、明天晴天(明天是晴天)

（6） For：为了/用于(目的关系)

N1 是 N2 的用途，包括目的和目标两个方面。

① 目的：公益项目（为了公益而...的项目）、慈善基金（为了慈善的而成立的基金）

② 目标：婴幼儿配方（用于婴幼儿的配方）、儿童牛奶（用来给儿童喝的牛奶）

（7） From(来源关系)

表示 N1 是 N2 事物的来源。

① 具体：中国游客（来自中国的游客），新华社消息（来自新华社的消息）

② 抽象：部门规定（来自部门的规定）、社会需求（来自社会的需求）

（8） Do：从事/教/生产(从事关系)

N1 是 N2 所从事的工作，可以是领域或者内容等。 N2 可以是人或者机构等。

① 领域：互联网精英（从事互联网领域工作的精英）、生物学者（从事生物相关研究的学者）

② 内容：钢铁工人（生产钢铁的工人）、英语教师（教英语的教师）

（9） Like(比喻关系)

N1 是对 N2 的比喻或者隐喻。往往可以用“像……一样”来解释。

例如： 金砖国家（像金砖一样的国家）、 影子银行（像影子一样的银行）

（10） Of(属性属于关系)

① 属于： 广汽丰田（属于广汽的丰田品牌）、集体财产（属于集体的财产）、消费者权益（属于消费者的权益）

② 属性： 中国特色（ 中国的特色）、西洋风格（西洋的风格）、产品质量（ 产品的质量）

（11） Locate/In(位于关系)

N1 和 N2 在空间上有包含和被包含的关系。

例如：杭州西湖、英国伦敦

（12） And(并列关系)

N1 和 N2 在语义上处于平等位置，属于并列列举关系。

例如：华人华侨、田间地头、爸爸妈妈

（13） Time(时间关系)

N1 和 N2 有时间上有先后或者时间点和时间段的包含关系。

例如：昨天下午、去年春天

（14） Content(内容关系)

N1 是 N2 所指称事物的具体内容。N2 通常是相对抽象和概括类的名词。

例如：能源项目（内容是能源方面的项目）、户籍政策（内容是有关户籍的政策）、质量报告（内容是有关质量的报告）

3.3 知识库构建

NN 短语知识库标注内容主要分为结构和语义两个方面。结构是指基本短语内部两个名词之间的结构关系；语义部分则包括三部分，一是组成短语的名词所属的语义类别，二是组成短语的名词之间的语义关系，三是基本复合名词短语是否指称一个命名实体。其中语义类别的标注根据上一小节中建立的语义关系体系进行标注。数据条目标注示例如表 1 所示：

表 1 数据条目标注示例

NN	语义类	结构	语义	实体	频次
记者 马晓	职业 人名	同位	Be 说明	1	233
木头 桌子	材料 用具	定中	Make 组成	0	651

3.3.1 短语结构关系标注

基本复合名词短语的内部语法结构包含四种：定中结构、联合结构、同位结构和主谓结构。以每种结构的拼音缩写为标记来对短语进行标记，即分别标记为 dz（定中）、lh（联合）、tw（同位）、zw（主谓）。

定中结构：定中短语是由有修饰关系的两部分组成，构成短语的两个名词中，前者为修饰语，后者为中心语的短语。例如:“中国人民”;“工作会议”。

联合结构：联合短语是由语法地位平等的两项组成，内部词语之间是联合关系的短语。例如：“华人华侨”;“好人好事”。

同位结构：同位短语是指构成短语的词语不同，但所指的是同一事物，且二者语法地位相同的短语。其与联合短语的不同之处在于，构成联合短语的两个名词指的是不同的事物，而同位短语则是用不同的词语指称相同的事物。例如：“首都北京”;“习近平主席”。

主谓结构：主谓短语是指由有陈述关系的两个成分组成的，前面是主语，表示说的是人或者事物，后面陈述的部分是谓语，说明主语的状态或者是什么的短语。例如：“今天星期一”;“明天晴天”。

3.3.2 名词语义类标注

本文采用北京大学的《现代汉语语义词典》（SKCC）为参考标准，进行短语内部名词语义类别的标注。根据 SKCC 的语义类体系，我们在其基础上添加了几个小类：一是在“个人”下面添加了“称谓”；二是将除人名、地名以外的命名词统一提出来标注为“命名词”，如建筑物名、游戏名、舞蹈名等。

对于名词的语义类标注，本文遵循以下几个原则：

① 标签尽量细致。也就是能标小类的尽量标注小类，实在无法细分的，可以标为上层标签。例如：“白色” ——颜色（抽象事物）；利益——抽象事物。

② 对于 SKCC 中没有的词，本文采用相似词标签相同的原则。例如：词典中没有“妇科”等词，但是有“骨科”，并且标签为“领域|处所”。根据词语相似原则，也将“妇科”标注为“领域|处所”。

③ 对于 SKCC 中同类词标签不一致的地方，要更正并标注正确标签。例如：词典中将“绿色/蓝色”标注为“颜色”，但把“红色”标注为“抽象事物”，把“紫红色”标注为“外形”。将颜色类的词统一标注为“颜色”。

④ 对用普通名词作为命名词的情形，“命名词”和原有的语义类都要标注，二者用“|”隔开。例如：“花园小区”中的“花园”，标为“命名词|建筑物”。

3.3.3 命名实体标注

对于狭义的命名实体，采用以下的判断原则：

① 符号性：命名实体是唯一一个个体的专有名称，是一个区别性称谓，具有代号性质。

② 命名实体所制成的事物不可向下再分类。

③ 命名实体所指称的事物是世上独一无二的，通常

不可以被数量词修饰；一旦以数量词修饰，命名实体就被转义了。

广义上，命名实体还包括一些非概念性称谓，如“厄尔尼诺现象”、“格力空调”。这种非概念性称谓一般只具备符号性，而不具备个体性和唯一性。在本文标注内容中，我们采用广义的命名实体定义，将基本复合名词短语中在位置上与命名实体紧密相连、语法词法上与命名实体紧密结合、语义上与命名实体概念范畴相同的词统一用数字标签“ 1”标注出来。

3.3.4 知识库建设成果

知识库的标注采用双盲标注，然后由第三人进行复核。短语句法标注一致率为 99.9%，短语语义关系标注一致率为 87.4%。最终形成了总共包含短语数 18281 条的 NN 短语句法语义知识库[3]。包括短语所包含的每一个名词的语义类、每一条短语的短语结构、语义关系和命名实体。形成了三个表，分别包含：1、短语和各个句法语义信息；2、语义类组合语义关系对应信息；3、语义类组合和是否为命名实体。知识库的最终标注成果规模如表 2 所示：

表 2　NN 短语知识库的规模

类别	类数	数量
短语	18281	18281
名词	7627	36562
语义类	96	37024
语义类组合	1133	18281
命名实体（单个）	2843	5886
命名实体（扩展）	3584	3584

4 NN 短语语义基本情况考察

4.1 分布情况

知识库前 50%的语义类出现的次数占到了所有语义类总数的 90%。出现频次排名前五的语义类有：抽象事物、地名、身份、处所及领域（图 1）。语义类组合较多，其中频次前五的语义类组合为：职业+人名、地名+地名、地名+处所、身份+人名、地名+机构。在排名最靠前的十个语义组合中（图 2），仍然以实体相关的名词组合为主，这说明基本复合名词短语在指称实体中应用相当广泛，这也离不开名词在人们在对事物的认知过程中的指称作用。而且作为报刊语料，往往需要指出事件发生的主体人物、地点以及相关机构等实体，因此其中出现的实体自然也会远多于其他语体的语料。

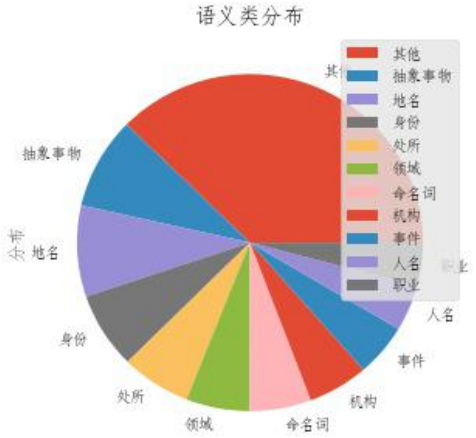

图 1 语义类分布

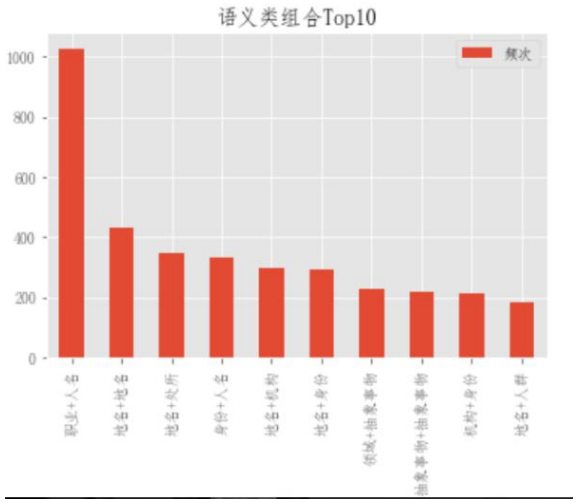

图 2 语义类组合 Top10

对十四类语义关系进行统计，其分布如表 3 及图 3 所示。

表 3　语义关系分布及对应典型语义类组合

语义	频次	典型语义类组合	示例
Of	5927	材料+量化属性	原油价格
Be	3399	人名+身份	习近平主席
Locate	3195	地名+处所	深圳市场

[3] 该库已经开源，共享在：
https://github.com/liupengyuan/Basic-Noun-Compounds

For	1411	抽象事物+事件	公益讲座
Content	992	抽象事物+ 抽象事物	能源项目
Do	750	领域+机构	金融研究所
Make	688	材料+非身体构件	铝合金轮毂
From	652	抽象事物+ 抽象事物	工作压力
Have	535	生理+身份	糖尿病患者
Use	303	材料+用具	汽油发动机
And	272	关系+关系	岳父岳母
Time	70	时间+时间	民国时期
Like	54	颜色+领域	绿色经济
Cause	32	生理+微生物	禽流感病毒

在这些语义关系中，出现频次最多的就是 Of/领属关系，这也符合我们对 NN 短语的认知。其中 Time[4]，Like 及 Cause 三种语义关系均在 100 以内，这是受到语料是新闻领域的影响，这三类语义关系的使用相对弱势。

图 3 语义关系分布

4.2 语义类组合与语义关系

相同语义类别的短语组成成分之间往往存在一定的联系，每一类语义关系通常对应多种语义类组合。表 3 展示了每一类语义关系中的典型语义组合并给出了示例。

理想情况下，我们希望语义类组合与语义关系是多对一对应的关系，但事实上，除了每一类语义关系对应多种语义组合之外，同一语义组合对应着多种语义关系。原因在于，语义类词典对名词分类的粒度还不够细，也就是说，出现此类情况意味着同一语义类的名词可能需要进行更细致的分类才能够进一步区分。由于语义类组合数目较多，因此表 4 仅给出了典型的语义类组合所对应的语义关系及示例。

表 4 典型的语义类组合所对应的语义关系及示例

语义类组合	语义关系	示例
材料+ 抽象事物	Of	甲醛含量
	Like	金砖国家
材料+ 创作物	Of	建材品牌
	Use	水墨作品
抽象事物+ 处所	Have	主题展区
	For	项目办公室
抽象事物+ 创作物	Have	共性问题
	Of	工作总结
处所+处所	Of	街道辖区
	Locate	内地城市
	Be	试点城市
处所+ 量化属性	Locate	全球销量
	Of	辖区面积
处所+领域	Locate	全国生态
	Of	世界历史
	For	太空科技
处所+身份	Locate	美国公民
	From	日本侵略者
	Of	印度总理

5 结语

本文以新闻媒体语料中抽取的基本复合名词短语为数据基础，结合语言学基本理论、已有复合名词短语语义关系分类及相关研究成果，建立了中文基本复合名词短语语义关系体系。在此基础上，对基本复合名词短语的语义类、语义关系、结构关系以及命名实体进行标注，建立了一个基本复合名词短语句法语义知识库。在该知识库的基础上，我们

[4] 类似“x 月 y 日”的短语，由于分词原因，本文没有统计在内。

对汉语基本复合名词短语的语义进行了初步分析。希望知识库能够为中文基本复合名词短语句法语义的研究提供基础数据支撑。

基本复合名词短语语义的人工标注存在一定的主观性，特别是语义关系。虽然采用双盲标注且一致率较高，但恐怕仍然无法符合所有人的语感。此外，由于知识库只选取了从新闻报刊语料中抽取的高频短语，类型较为单一，平衡性较差，因此，对于语义关系的分类和描述仍然有待进一步完善。下一步工作的重点将针对语料的平衡性和人工标注的一致性，对整个语义关系体系进行调整、细化或修改，选取多个领域的数据，进行标注，争取建设成为平衡性、规模和质量三者均衡的中文基本复合名词句法语义知识库。

参考文献

[1] Downing P. On the creation and use of English compound nouns[J]. Language, 1977: 810-842.

[2] 吕叔湘.现代汉语语法提纲 (油印本)，1976.

[3] Vanderwende L. Algorithm for automatic interpretation of noun sequences[C]. Proceedings of the 15th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 1994: 782-788.

[4] Vanderwende L. SENS: the system for evaluating noun sequences[M]. Natural language processing: The PLNLP approach. Springer US, 1993: 161-173.

[5] Tratz S, Hovy E. A taxonomy, dataset, and classifier for automatic noun compound interpretation[C]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2010: 678-687.

[6] 魏雪,袁毓林. 基于规则的汉语名名组合的自动释义研究[J]. 中文信息学报, 2014, 28(3):1-10.

[7] 魏雪,袁毓林. 基于语义类和物性角色建构名名组合的释义模板[J]. 世界汉语教学, 2013(2):172-181.

[8] Levi J N. On the alleged idiosyncrasy of non-predicate NP's[C].Chicago Linguistic Society. 1974: 10.402-15.

[9] Levi J N. The syntax and semantics of complex nominals[M]. Academic Press, 1978.

[10] Warren, B. Semantic patterns of noun - noun compounds. Gothenburg: Gothenburg University Press. 1978.

[11] Barker, K. and S. Szpakowicz. Semi-Automatic Recognition of Noun Modifier Relationships. In Proc. of the 17th International Conference on Computational Linguistics.1998.

[12] Nastase V. and S. Szpakowicz. Exploring nounmodifier semantic relations. In Proc. the 5th International Workshop on Computational Semantics.2003.

[13] Girju, R. 2007. Improving the interpretation of noun phrases with cross-linguistic information. In Proc. of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007).

[14] D. Ó Séaghdha and A. Copestake, "Co-occurrence Contexts for Noun Compound Interpretation," Proc. ACL-07 Workshop Broader Perspective Multiword Expressions, Prague, Czech Republic, 2007.

[15] 袁毓林. 谓词隐含及其句法后果——"的"字结构的称代规则和"的"的语法、语义功能[J].

中国语文,1995(04):241-255.

[16] 张卫国,梁社会.定语类型和槽关系类型的对应及其对名词语义分析的作用[C]// 全国计算语言学联合学术会议. 2003.

[17] 李宇明. 非谓形容词的词类地位[J]. 中国语文,1996(01):1-9.

[18] 文贞惠. “N_1(的)N_2”偏正结构中N_1与N_2之间语义关系的鉴定[J]. 语文研究,1999(03):22-27.

[19] 蔺璜. 定语位置上名词的句法表现及其语义特征[J]. 山西大学学报哲学社会科学版, 2005, 28(2):95-99.

[20] 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(04):342-355+384.

[21] 黄国营.“的”字的句法、语义功能[J].语言研究,1982(01):101-129.

[22] 孔令达. “名_1+的+名_2”结构中心名词省略的语义规则[J]. 安徽师大学报(哲学社会科学版),1992,(01):103-107.

[23] 单强,牛守祯. 定语的语义类型举要[J]. 潍坊教育学院学报,1999,(04):27-30.

[24] 马洪海.“名+名”组合的语义考察[J].信阳师范学院学报(哲学社会科学版),1999(01):117-120.

[25] 周日安. 名名组合的句法语义研究[D].暨南大学, 2007.

[26] 魏雪. 面向语义搜索的汉语名名组合的自动释义研究[D]. 北京大学, 2012.

[27] 鲁川.智能计算机的知识表示和汉语的语义研究[J].语文建设,1992(11):30-33.

[28] 林杏光, 张庆旭. 现代汉语槽关系研究[J]. 汉语学习, 1998(6):7-10.

[29] 冯志伟.中文信息MMT模型[J].语言文字应用,1992(04):21-30.

[30] 董振东,董强,郝长伶.知网的理论发现[J].中文信息学报,2007(04):3-9.

[31] 刘开瑛. 汉语框架语义知识库构建工程[A].中国中文信息学会.中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集[C].中国中文信息学会:,2006:8.

[32] B. Rosario and M. Hearst, “Classifying the Semantic Relations in Noun Compounds via a Domain-Specific Lexical Hierarchy,” Proc. Conf. Empirical Methods Natural Language Process.,Somerset, NJ, USA, 2001, pp. 82-90.

[33] Kim, S.N. and T. Baldwin. Automatic Interpretation of Compound Nouns using WordNet::Similarity. In Proc. of 2nd International Joint Conf. on Natural Language Processing.2005.

[34] 肖国政. 基于语义依存图的汉语复杂名词短语资源建设与自动分析研究，国家自然科学基金项目61173095. 2011.

[35] 卢涌.面向深层语义表示的“名词+的+名词”结构语义释义研究[D],北京语言大学，2017.

刘鹏远，

地址：北京语言大学信息科学学院，

邮编：100083，

电话：13911510796，

E-mail: liupengyuan@pku.edu.cn