汉语复合名词短语语义关系知识库构建与自动识别研究*

张文敏[1]，李华勇[1]，邵艳秋[1]

（1.北京语言大学 信息科学学院，北京 100083）

摘 要：汉语复合名词短语因其使用范围广泛、结构独特、内部语义复杂的特点，一直是语言学分析和中文信息处理领域的重要研究对象。国内关于复合名词短语的语言资源极其匮乏，且现有知识库只研究名名复合形式的短语，包含动词的复合名词短语的知识库构建仍处于空白阶段，同时现有的复合名词短语知识库大部分脱离了语境，没有句子级别的信息。针对这一现状，该文从多个领域搜集语料，建立了一套新的语义关系体系，标注构建了一个具有相当规模的带有句子信息的复合名词语义关系知识库。该库的标注重点是标注句子中复合名词短语的边界以及短语内部成分之间的语义关系，总共收录 27007 条句子。该文对标注后的知识库做了详细的计量统计分析。最后基于标注得到的知识库，该文使用基线模型对复合名词短语进行了自动定界和语义分类实验，并对实验结果和未来可能的改进方向做了总结分析。

关键词：汉语复合名词短语；语义关系体系；定界识别

中图分类号：TP391　　**文献标识码：**A

Chinese Compound Noun Phrases Semantic Relations Knowledge Base Construction and Auto Recognition

ZHANG Wenmin[1]，LI Huayong[1]，SHAO Yanqiu[1]

(1.Information Science School, Beijing Language and Culture University, Beijing 100083,China)

Abstract: Chinese compound noun phrases are characterized by their wide range of use, unique syntactic structure and complex internal semantics, which has always been an important research object in the field of linguistic analysis and Chinese information processing. The language resources of compound noun phrases are extremely scarce in China, and the existing knowledge base only studies noun-compound phrases, while the construction of a knowledge base containing compound noun phrases with verbs is still in the blank stage. At the same time, most of the existing knowledge bases of compound noun phrases are out of context and have no information at sentence level. In accordance with the present condition, this paper collects corpus from many fields, a new semantic relation system is established. In addition, a compound noun semantic relation knowledge base with sentence information is constructed by annotation. The focus of the library is to mark the boundary of compound noun phrases in sentences and the semantic relationship between the internal components of the phrases. A total of 27007 sentences are collected. This paper makes a detailed statistical analysis of the annotated knowledge base. Finally, based on the annotated knowledge base, this paper uses the baseline model to carry out automatic delimitation and semantic classification experiments for compound noun phrases, and summarizes the experimental results and possible improvement directions in the future.

Key words: Chinese compound noun phrases; Semantic Relational System; Delimitation recognition

0. 引言

复合名词短语在日常生活中应用广泛，在语言使用中占有较大比重。据 Leonard[1]统计，近两个世纪以来，在小说体散文中使用名词复合短语的次数呈现稳定持续增长的态势，同时名词复合短语的种类也有显著的增长。且其语法结构较为独特,语义关系较为复杂，因此在语言分析中扮演着非常重要的角色，通过对它的定界识别和语义分类可以有效改善句子语义分析的质量，进行信息的准确抽取。

对于复合名词短语的研究，国外很早就有相关的语义关系体系的建设研究，关于短语的边界识别和语义分析，也相对于国内而言较为成熟，大概有[2-9]等。

国内以往对汉语基本名词短语的研究，主要是基于边界识别和自动释义，而针对短语内部构成成分之间的语义关系体系建设却相对较少，目前较完整的是刘鹏远[10]针对名名复合形式短语语义知识库的构建，但该文只是单纯从语言学角度进行了语义分类，做了一些初步的统计分析，并没有将包含动词的复合名词短语纳入研究范围，且抽取得到的复合名词短语脱离语境，缺少句子级别的信息。目前也没有在句子中进行复合名词短语自动定界和语义关系分类的研究工作。

针对国内对于复合名词短语语义知识库构建相对薄弱的特点,我们参照北京大学《现代汉语语义词典》的语义类别标签并结合语料的实际情况建立一个语义

*收稿日期：　　定稿日期：

通讯作者：邵艳秋

基金项目：国家自然科学基金项目（61872402）；教育部人文社科规划基金项目(17YJAZH068)；北京语言大学校级项目（中央高校基本科研业务费专项资金）（18ZDJ03）

作者简介：张文敏（1993—），女，硕士研究，计算语言学；李华勇（1994—），男，硕士研究生，计算语言学；邵艳秋（1970—），女，教授，计算语言学。

关系体系，标注构建了一个包含句子信息的复合名词短语语义关系知识库，短语的构成成分包括名词和动词。基于知识库，对语义关系类型的分布情况和词性分布的特点做了统计分析。最后基于此知识库，我们了构建相应的数据集，采用 BERT+Bi-LSTM+CRF 模型做了定界识别和语义关系分类的实验。针对实验结果进行了分析总结，并讨论了未来可能的改进方向。

本文后续组织如下：第 1 节对以往相关研究工作进行综述；第 2 节解释语义关系体系内容；第 3 节介绍知识库的基本情况，包括语料的来源，标注过程，统计分析等；第 4 节介绍实验模型的相关情况；最后一节对全文进行总结。

1. 相关研究

1.1. 复合名词短语语义关系体系研究和知识库建设现状

国内外关于复合名词短语语义关系的研究主要采用两种方法，一种是通过整理总结复合名词短语内部各成分之间的语义关系类来定义其语义关系，另一种是基于谓词语义类来确定复合名词短语内部成分的语义关系。

国外研究中，Downing[11]针对英语复合名词短语提出了十二类语义关系；Levi[12,13]通过删除谓词，提出了十二类名名复合名词短语成分之间的语义关系；Warren[14]认为复合名词短语的语义关系由四个层级组成，最顶层有六类粗粒度语义关系，各个粗粒度关系下又分为其他细粒度的关系类型；2007 年 SemEval[15]组织了一项评测“Classification of Sematic Relations between Nominals”，定义了七种语义关系。

Tratz&Hovy[16]建立了目前最大的英语复合名词短语语义关系知识库，含 17509 条短语，十二类语义关系，每一类关系下又分了小类，并做了关系标注。

汉语方面，马洪海[17]考察“名+名”组合，偏正结构分为七类语义关系，复指结构分为八类；魏雪[18]针对汉语复合名词短语归纳出 26 种语义组合关系；Jinglei Zhao[19]参考动词的语义角色为 300 个名词短语标注了四种粗粒度语义关系；刘鹏远[10]为名名组合的复合名词短语定义了十四种语义关系。

关于知识库的构建，目前有魏雪和袁梳林[18，20]以识别隐含谓词和自动释义为目的而建立的名名搭配知识库，但是该知识库目前尚未开源。还有刘鹏远[10]对 18281 条名名复合名词短语进行标注而形成的知识库，该知识库不仅标注了两个名词之间的语义关系，同时也标注了两个名词各自的语义类，但语义类组合和语义关系呈现多对多的情况，严重影响了数据分析，若要解决此问题，又需进一步进行更细的名词语义分类，这无非对后续工作增加了更大的难度和人力投入。

1.2. 复合名词短语定界识别研究现状

早期国内对于名词性短语的边界识别研究经常与语法分析联系在一起，多使用基于统计的方法。赵军[21]将表示 baseNP 内部句法组成结构模板与体现上下文约束条件的 N 元模型结合起来形成一个新的模型，识别结果准确性明显优于单纯基于词性标注的 N 元模型，但不足之处是对上下文句法特征不明显的 baseNP 识别精确率较低；孟迎[22]从语料库中自动抽取基本名词短语的词性模板及其相应的上下文信息并采用算法形成相应的决策树来识别汉语名词短语。但以上实验都着重于名词性短语句法结构的研究，对短语内部语义关系也没有作深入的探讨，且研究对象基本未包含动词在内。

祝慧佳[23]采用三种方法对长度在 2-10 不等的复合名词短语进行边界识别，所研究的短语对象虽包含了动词，但只是针对动词采取词性细分类标注来提高识别准确率；孙玉祥[24]提出一种融合统计机器学习与后处理规则相混合的识别策略，但后续的处理规则同样未涉及短语内部成分之间的语义特征进行识别。还有其他相关研究主要采用统计和规则两种方法进行边界识别，并没有将语义关系应用于边界识别的任务当中来。

2. 语义关系体系

2.1. 复合名词短语具体概念分析

在汉语词汇研究中，赵军[21]从限定性定语出发对汉语基本名词短语（BNP）进行了形式化的定义：

Base NP → Base NP + Base NP

Base NP → Base NP + 名词|名动词

Base NP → 限定性定语 + Base NP

Base NP → 限定性定语 + 名词|名动词

限定性定语 → 形容词|区别词|动词|名词|处所词|西文字串|（数词+量词）

以往针对汉语名词性短语的研究工作大都基于此概念，另外还有孙玉祥[24]基于基本名词短语和最长名词短语提出的简单名词短语（SNP），SNP 按其结构特点分为七类，内部不仅包含复杂的并列或嵌套结构，还包含动宾和专有名词性结构，粒度大于 BNP，因此我们选择以赵军定义的基本名词短语为准。

由于汉语属于意合语言，名词性短语在构成上较为灵活，构成成分通过简单的组合就可以构成短语序列，不需要助词等连接成分，如“医疗设备、国际经济政治”；且动词作为复合名词短语的构成成分时没有变形信息，它既可充当名词的功能作短语的核心词，如“工作接洽、多边会谈”，又可以动词的成分作短语的修饰语或修饰语的一部分，如“行军路线、安全管

理办法”。由两个实词构成的复合名词短语在整个名词性短语中的占比大，且由于形容词参与构成复合名词短语时其语义功能和语法组合类型都较为单一，语义分类和边界识别任务相对来说较为简单，因此本文的研究对象为由动词和名词参与构成的长度为 2 的复合名词短语。

关于“V+N”中“V”是否可以被划分为名动词，陆俭明[25]认为能直接作定语的动词只是动词的一个小类，叫名动词；邵敬敏[26]认为如果将直接修饰 N 的 V 认定是名动词，那么名动词的范围会无限扩大，需要限制其范围；尹世超[27]以动词是否能直接作定语，将动词分为可定动词和不可定动词。虽然语言学家各有论断，但是都认为动词处于修饰语位置上时其本身的性质有所变化，有必要划分出名动词的类别，所以，在本文以下表示中，我们将处于限定词位置上的动词标明为名动词，而将被修饰的动词还是标明为动词。

基于赵军的基本名词短语定义，我们将本文所研究的复合名词短语形式化表示为：

复合名词短语 = 限定词+核心词

限定词 → 名词|名动词

核心词 → 名词|动词

其中可能出现的组合形式有：名词+名词、名词+动词、名动词+名词三大类，其中名词包含有普通名词、专有名词（人名、地名、机构名、品牌名）、时间名词、处所名词等。

2.2. 复合名词短语的语义关系类型

本文采用的语义关系标注规范参考北京大学的《现代汉语语义词典》中的语义分类标签，并结合本文具体任务作了一些调整，最后确定十种语义关系，即时间、处所、领域、名称、材料、并列、式样、用途、内容、一般修饰。

由于复合名词短语的构成成分有名词和动词，且二者排列次序不等，所以在此我们以“词 1+词 2”的形式表示短语。

（1）时间

词 1 是时间名词，表明的词 2 所处的时间状态或具有的时间属性。

eg1:他们的谈话，若能记录下来，一定是历史学家极感兴趣的中国**近代城乡**的变迁史料。

说明：在近代形成的城乡

eg2:一个**成年男子**看着一个小孩在小溪里玩耍。

说明：处于成年时段的男子

（2）处所

词 1 是处所名词，表明词 2 所处的空间地理位置。

eg1:四名年轻女子围在**厨房柜台**前，面前摆着一盘布朗尼蛋糕。

说明：摆放在厨房的柜台

eg2:古代亚历山大的事件预计会对**港口活动**产生重大影响。

说明：在港口举办的活动

（3）名称

词 1 是专有名词，包括人名、地名、国名、品牌名、机构名等，处于限定语的位置交代了词 2 的国别、品牌、称谓等信息。

eg1:我感觉**卡文迪许太太**把它藏起来了。

eg3:一个穿着红色**阿迪达斯运动衫**，戴着红色太阳镜，戴着红色帽子的男人穿过小镇。

（4）式样

词 1 表示词 2 的款式、颜色、形状、架构等外部特征或表面形态。

eg1:一个穿着**条纹衬衫**的男孩牵着一条小狗。

说明：衬衫的表面图案呈条纹状

eg2:一个女人在小亭子上装饰着**杯形蛋糕**，旁边的人仔细地观察着这个技巧。

说明：外观形状像杯子状的蛋糕

eg3:旁观者从**露天看台**观看时，一匹马在竞技场上抢走了它的骑手。

说明：看台的建筑架构是露天无顶的

（5）材料

词 1 是构成词 2 的原材料，词 2 一般表示人工制成品。

eg1:如果粗糙的纺织品不吸引人，那么你也可以找到漂亮的刺绣品，比如棉布、**亚麻桌布**和餐巾。

说明：用亚麻编织成的桌布

eg2:一个穿牛仔夹克的男人走过一个华丽的**石头拱门**。

说明：由石头堆砌成的拱形门

（6）并列

词 1 和词 2 的语义信息平行，语法地位等同，组合构成并列。

eg1:一个**金发碧眼**的女人在俱乐部唱歌。

Eg2:**街坊邻居**现在最常一起做的娱乐就是到俱乐部来运动。

（7）用途

词 1 表示词 2 产生的目的，即作何用处。

eg1:印第安纳州正在研究文件**汇编软件**，伊利诺伊州正在研究音频视频会议与文件汇编的结合。

说明：用来对文件进行汇编的软件

eg2:一个拿着**购物袋**的女人从地铁旁走过。

说明：购物时使用的专用袋子

（8）领域

词 2 通常是较为抽象或概括性较强的词语，词 1 表示词 2 领域范围中的一类或对词 2 的具体化解释说明。

eg1:一群**足球运动员**正在踢足球。

说明：运动项目是踢足球的运动会

eg2:四名男性**建筑工人**站在一起，其中三名身穿

黄色衬衫。

说明：从事建筑行业的工人

（9）内容

一般词 2 具有容载性，而词 1 表示词 2 的内容或词 2 所包含传达的信息。

eg1:95%的当地新闻报道犯罪和**灾难画面**，5%是可爱的动物片段。

说明：呈现内容是灾难场景的画面

eg2:她后来的表现打破了传统党外女性的**参政经验**，一改鲜明的受难者家属形象，展现了女性政治人物的主体性。

说明：有关于参政的经验

（10）一般修饰

这一语义关系类别包括以上语义关系类型之外的其他所有可能类型，但主要是词 1 表示词 2 的属性、类型或领属，但也包含其他类别。

例如：

① 属性：

eg1:畜牧业的过度发展还使**大片草原**变成沙漠。

说明：大片的草原

eg2:审核员应使用他们的**专业判断**，来确定沟通的形式和内容。

说明：专业的判断

② 类型：

eg1:两名男子玩**电子游戏**。

说明：电子类的游戏

eg2:最近国际发布了全球三十八个国家的国中生自然科学和**数学成绩**报告。

说明：数学科目的成绩

③ 领属:

eg1:不但如此，肉食吃多了，**动物脂肪**会使血管渐渐失去弹性，久而久之极易引起动脉硬化，从而诱发高血压和心脏病。

说明：动物体内的脂肪

eg2:每个人用左手按住饭盆或菜盆的边儿，用**右手手指**抓自己面前的饭和菜，放入口中。

说明：右手包括手指

④ 其他：

eg1:世平觉得**单身女子**需要这样的设备。

说明：处于单身状态的女子

eg2:与其称之为乐团，不如将她们看做美少女**偶像团体**。

说明：走偶像路线的团体组合

3. 知识库建设

基于 2.2 中定义的语义关系体系，我们建立了一个语义关系知识库。不同于现有知识库，我们的知识库同时提供句子和句子中复合名词短语的边界以及语义关系。构建知识库需要先收集大量的多领域句子，再经过数据清洗和预筛选，得到待标注数据，然后借助标注平台，由标注员进行标注。标注员首先需要标识出复合名词短语在句子中的位置，然后对其做语义关系分类。与此同时我们借助标注平台对标注的质量和一致性进行监督。整个知识库的构建过程可以分为：（1）生语料收集和预筛选，（2）组织标注，（3）语料统计分析。

3.1. 生语料收集与预筛选

为了使最后的知识库在有限数据量的情况下尽可能包含各种自然语言现象，同时体现语义关系的真实占比，我们收集了多个不同领域的无标注数据，分别来自：新闻、论坛、现代小说、现代散文、剧本、对外汉语教材、中小学语文课本等 7 个不同领域。基于标点符号，对所有文本进行句子切分，筛除长度过长（超过 100 个字）和过短（不足 10 个字）的句子，最后得到约 10 万句生语料。

由于生语料的规模比较大，为了提高知识库建设的速度，减轻标注员的工作量，我们需要对生语料做预筛选，尽可能排除不包含复合名词短语的语料。根据本文定义的复合名词短语，限定词和核心词在语义依存分析结果中应当存在依存弧，又因为限定词和核心词主要为名词和动词，因此我们可以基于词性标签和依存弧对文本做进一步筛选过滤，最终我们筛选得到了 4 万 5 千句左右待标注的句子。

3.2. 标注过程

复合名词短语知识库的标注过程分为两个子任务，一是在句中确定有无复合名词短语，如果有则需要标识出复合名词短语的边界，如果没有则标“无 NP”；二是对标识出的复合名词短语做语义关系分类。

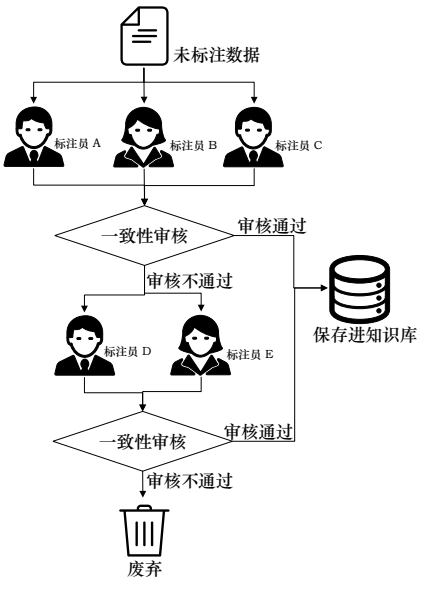

图 1 标注过程图示

为了方便标注过程，我们开发了一个 Web 标注工具，同时组织五名语言学硕士进行标注工作。在正式标注前，对标注员进行了为期两天的培训，每个标注

员试标 500 句，然后根据标注结果再进行统一修正。

标注过程如图 1 所示。首先，将每一句待标注文本分别发送给 A，B，C 三位标注员，三位标注员独立完成所有标注后系统会自动计算结果的一致性，如果一致性大于或等于 85%，则认为标注结果可靠，此时会随机抽取一个人的标注结果作为最终标注结果，保存进知识库。如果一致性小于 85%，系统会将该文本自动发送给 D，E 两位标注员，同时舍弃 A，B，C 的标注结果，由 D，E 做第二轮标注，然后系统计算第二轮的标注一致性，如果一致性大于 90%，则认为标注可靠，此时会随机抽取一个人的标注结果作为最终结果，存入知识库。如果第二轮的标注一致性小于 90%，则舍弃该文本。

在整个标注过程中，标注平台会自动为标注员动态分配标注身份，标注员不知道自己处在第几轮标注中，也无法看到其他标注员的标注结果，这样就保证标注过程互不干扰，同时确保了标注一致性的可信度。

最终，我们得到了 27007 条有效标注句子，整体复合名词边界一致性为 96%，复合名词语义关系一致性为 87%。

3.3. 语料统计分析

标注完成之后，我们对知识库进行了基本的统计分析。语料来源的分布情况如图 2 所示：

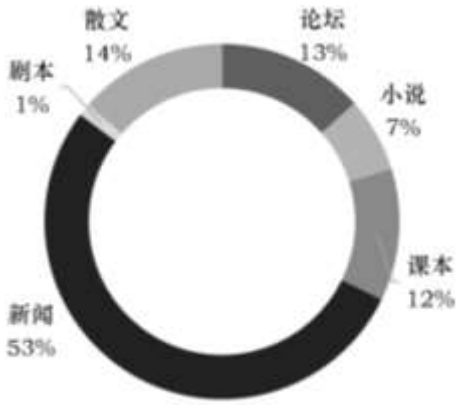

图 2 复合名词短语语料来源分布

语义关系分布如图 3，排名靠前的语义关系分别有:一般修饰、内容、名称。一方面是因为我们的语料一半以上来源于新闻领域，新闻用语较为正式端庄，构成名词性短语的两个成分呈领属关系的可能性较大，而领属关系包含在我们所定义的一般修饰关系当中，另一方面是因为一般修饰关系下位关系类型较为错综复杂，构成成分比较多，因此构成复合名词的两个词之间呈一般修饰关系较为普遍；其次占比较多的是内容关系类型，这说明名词性短语的第二个词表抽象概括性的居多，而第一个词起缩小第二个词范围的作用或具体第二个词所指事物的领域；名词的属性就是具有指称性，所有表名称义的词基本都是名词，包括人名、地名、品牌名、行政单位名称等，所以复合名词短语的语义关系中表名称语义关系的占比自然也是排在前位的。

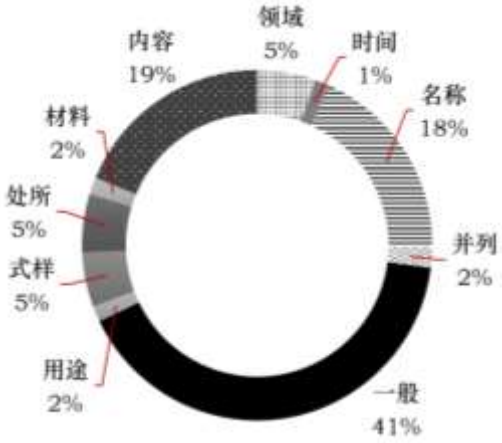

图 3 复合名词短语语义关系分布

我们对所有复合名词短语的词性组合分布进行了统计，如表 1，数据结果和语义关系的分布占比结果具有一致性，两个普通名词进行组合的数量最多，大部分情况下构成了一般修饰关系；名称名词和基本名词组合构成名称语义关系，排第三位。

表 1 复合名词短语的词性组合统计

词性组合	频次	说明
n + *n*	15358	基本名词+基本名词
v + *n*	4723	名动词+基本名词
nh + *n*	3300	名称名词+基本名词
n + *v*	1399	基本名词+动词
ns + *n*	1109	处所名词+基本名词
nz + *n*	679	品牌名词+基本名词
nt + *n*	354	时间名词+基本名词
nl + *n*	231	处所名词+基本名词
n + *nh*	196	基本名词+人称名词
ns + *ns*	183	处所名词+处所名词

由上可以初步推断内容语义关系的复合名词短语其构成成分多包含动词，因此我们进一步统计了名动词和动词在各个语义关系中的出现频次，如图 4 和图 5。动词在一般修饰关系中出现最多，说明名词位于动词之前主要就是起修饰限定的作用，比如“国民储蓄、常规表演、商务旅行”等，语法上这些组合中的第二个词都是动词，语义上属于一般修饰关系中的被修饰成分；排第二位是领域，是因为动词“比赛”在领域关系标签中出现次数较多；动词在内容语义关系中的出现频次也就较多，说明对于部分名词性“N+V”短

语，N 是 V 的受事、对象。因此，当名词位于动词之前构成一个名词性短语时，名词对动词的语义特征主要有修饰限定、领域分类、受事对象。

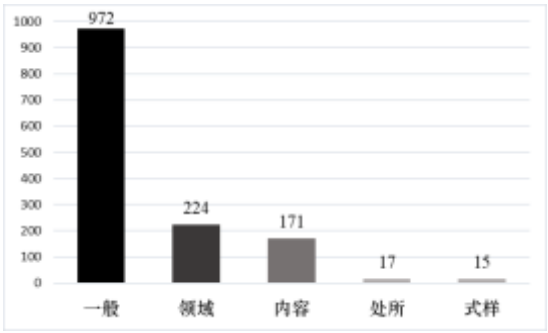

图 4 动词在语义关系中的分布情况

名动词的动作性较弱，具有名词的某些特点，一般叙述的是某一类事物，可以被解释为“关于 V 的 N”，因此基本表示的是关于名词的某些内容，其次还有部分名动词表示事物稳固的功能属性，例如“实验设备、分析方法”等，因此在用途语义关系中也有出现。

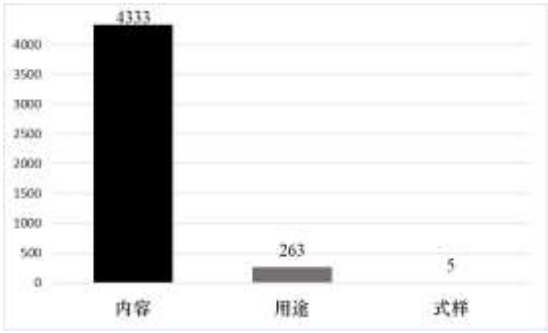

图 5 名动词在语义关系中的分布情况

4. 自动定界和语义分类研究

为了进一步研究该知识库对自然语言处理任务的帮助，我们初步尝试了基于知识库对复合名词短语进行自动定界和自动语义分类的任务。由于该知识库中不同语义关系的数据量差异较大，同时复合名词的语义分类需要较多的语言学知识，因此自动定界和语义分类任务具有一定的挑战性。

4.1. 任务定义与数据集划分

我们将复合名词短语的定界和语义分类建模为一个序列标注任务[28]。对于输入句子$X = x_1, x_2, x_3, \ldots, x_n$，模型需要为序列中的每个词（或者字）预测出对应的标签$Y = y_1, y_2, y_3, \ldots, y_n$，其中$y_i \in \{B, I, O\}$。BI 标签同时带有语义关系分类标签。这样，我们就将复合名词定界与语义关系分类组合为一个序列标注任务。同时，我们将标注后的知识库导出为序列标注格式文件，采用 BIO 标注体系。然后随机打乱顺序，划分为训练集，验证集和测试集。整个数据集的统计结果如下表所示：

表 2 数据集基本信息

数据集	句子数
训练集	18906
验证集	5401
测试集	2700

4.2. 基线模型

我们选择基于上下文语境词向量 BERT+双向 LSTM+CRF[29]的模型作为实验的强基线模型，如图 6 所示，整个模型包含三个部分：BERT 编码层，双向 LSTM 表示层，CRF 解码层。

我们使用 Google 开源的中文字符级预训练 BERT 模型，使其首先在超大规模的语料上进行预训练，得到良好的语义表示能力之后再将其接入到下游任务中充当表示层或者编码层。不同于传统的 Word2Vec 或者 Glove 词向量模型，BERT 输出的词（字）向量考虑了句子的语境，能够更好地表示词（字）的多义现象和语境信息。

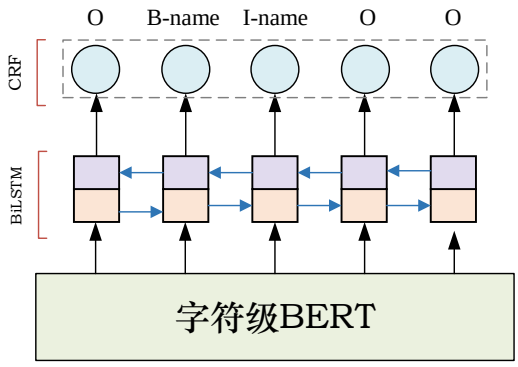

图 6 基线模型网络示意图

之后我们连接一层双向 LSTM 作为深度表示层，通过复合名词定界和语义分类任务的训练，表示层能够从 BERT 的丰富表示中有效抽取对我们任务真正有效的信息，同时舍弃不必要的干扰信息。最后我们接入一层 CRF 解码层，CRF 能够建模条件概率$P(y|x)$，在解码时，CRF 能够利用上下文信息作为特征，同时执行全局归一化，能够更好地预测标签序列。

4.3. 实验结果与分析

从表 3 可以看出，我们的模型整体识别能力仍有很大提升空间，大部分类别的 F1 得分都较低，最高值为“式样”语义关系，最低值为“并列”语义关系。从召回率和精准率上看，大部分语义关系的召回率都明显低于精准率，说明模型在识别正例的时候过于严格。根据数据集的特点，我们认为现有模型的问题主要有：

- 在一层 CRF 中同时解码复合名词的边界和语

义关系，难度较大

- 对于不同语义关系的区分缺少背景知识，由模型直接做 10 分类难度很大
- 数据集分布不平衡，部分语义关系的数据较少，模型难以学习到差别

表 3 实验结果

语义关系	精准率（P）	召回率（R）	F1 得分	训练集数量
处所	79.23%	58.98%	67.62%	1002
内容	71.34%	47.03%	56.69%	3451
一般	58.92%	51.97%	55.23%	7677
名称	61.55%	58.33%	59.90%	3463
领域	65.53%	52.12%	58.06%	1036
用途	72.24%	61.89%	66.67%	286
式样	71.66%	83.40%	**77.09%**	940
并列	52.04%	53.80%	52.91%	355
材料	72.61%	73.79%	73.19%	309
时间	58.78%	67.18%	62.70%	259
平均	**66.39%**	**60.85%**	**63.01%**	**1878**

基于强基线模型的结果和错误分析，我们认为，复合名词短语的定界和语义分类是一项具有一定挑战性的任务，未来的模型尝试可以考虑如下几个方向：

- 拆分定界任务和语义关系识别任务，采用多任务模型联合学习
- 引入语言学背景知识，提升模型的语义分类能力
- 基于伪数据增强的方式，缓解数据集的不平衡问题
- 基于 few-shot 学习的方式，缓解少样本下的学习困难问题

5. 结语

本文从来自多个领域的句子中标识包含动词的复合名词短语，基于北京大学《现代汉语语义词典》的语义类修改建立了复合名词短语的语义关系体系，对标识出的短语进行语义关系标注，构建一个语义知识库。基于该知识库做了词性和语义类型的统计分析，并用 BERT 和双向 LSTM+CRE 的强基线模型对复合名词短语进行定界和语义分类。希望为以后复合名词短语语义关系的研究提供语言资源方面的支持，为以后对复合名词短语的定界识别和自动语义分类提供帮助。

语义关系体系中，一般修饰的包含成分较为复杂多样，导致不同语义关系类的数据差异性明显，直接影响了后期的模型试验结果，所以其下位关系还需进一步探讨研究。此外，动词作为复合名词短语的构成成分，其自身的语义特征对短语内部成分的语义关系具有非常重要的意义，我们还需尽量多的收集包含动词的名词性短语，逐步完善各种组合形式的复合名词短语研究。下一步工作的重点是对一般修饰类的语义关系进行进一步的梳理切分，逐步完善语义关系类别，对语料来源再扩大范围，尽量使知识库中的复合名词短语更具代表性。

参考文献

[1] Leonard and Rosemary. The interpretation of English noun sequences on the computer[M]. North-Holland, 1984: 429.

[2] Nakov P, Hearst M. Search engine statistics beyond the n-gram: application to noun compound bracketing[C]// Conference on Computational Natural Language Learning, 2005: 17~24.

[3] Lauer M . Designing Statistical Language Learners: Experiments on Noun Compounds[J]. Computer Science, 2012.

[4] Kim S N, Baldwin T. Automatic Interpretation of Noun Compounds Using WordNet Similarity[C]// International Joint Conference on Natural Language Processing. 2005: 945~956.

[5] Lapata M. The Disambiguation of Nominal isations[J]. Computational Linguistics, 2002, 28(3):357-388.

[6] Moldovan D , Badulescu A , Tatu M , et al. Models for the semantic classification of noun phrases[J]. In HLT-NAACL 2004: Workshop on Computational Lexical Semantics, 2004:60--67.

[7] Vanderwende L . Algorithm For Automatic Interpretation Of Noun Sequences[J]. Proceedings of COLING-94, 1994: 782~788.

[8] Barker K, Szpakowicz S. Semi-automatic recognition of noun modifier relationships[C]// International Conference on Computational Linguistics. 1998: 96~102..

[9] Rosario B, Marti H. Classifying the Semantic Relations in Noun Compounds via a Domain Specific Lexical Hierarchy. In: Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. 2001: 82~90.

[10] 刘鹏远,刘玉洁. 中文基本复合名词短语语义关系体系及知识库构建[J]. 中文信息学报, 2019.

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

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

[13] Levi JN. The syntax and semantics of complex nominals[M]. Academic Press, 1978.

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

[15] Diarmuid Ó Séaghdha. SemEval-2010 Task 9: The

Interpretation of Noun Compounds Using Paraphrasing Verbs and Prepositions[C]// Workshop on Semantic Evaluations: Recent Achievements & Future Directions. Association for Computational Linguistics, 2010.

[16] Tratz S , Hovy E H . A Taxonomy, Dataset, and Classifier for Automatic Noun Compound Interpretation[C]// Acl, Meeting of the Association for Computational Linguistics, July, Uppsala, Sweden. DBLP, 2010: 678-687.

[17] 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(1):117-120.

[18] 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.

[19] Zhao, Jing lei, Hui Liu and Ruzhan Lu. Semantic Labeling of Compound Nominalization in Chinese[C]// Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, June 2007 :73-80.

[20] 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.

[21] 赵军,黄昌宁.结合句法组成模板识别汉语基本名词短语的概率模型[J].计算机研究与发展,1999,(11).

[22] 孟迎,冯丽辉等.基于决策树的汉语基本名词短语识别[J].黑龙江工程学院学报(自然科学版),2004,(6).

[23] 祝慧佳. 汉语名词复合短语识别与分类的方法研究[D].哈尔滨：哈尔滨工业大学,2007.

[24] 孙玉祥. 汉语简单名词短语自动识别的研究［D］. 大连:大连理工大学，2014.

[25] 陆俭明. 汉语和汉语研究十五讲[M]. 北京: 北京大学出版社,2004.

[26] 邵敬敏. 双音节结构的配价分析[J]. 现代汉语配价语法研究. 北京:北京大学出版社,1995.尹世超.动词直接做定语与动词的类[A]. 第十一次现代汉语语法学术讨论会.2000.

[27] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C] Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 1064-1074.

[28] Yang, Jie, Shuailong Liang, and Yue Zhang. "Design Challenges and Misconceptions in Neural Sequence Labeling." Proceedings of the 27th International Conference on Computational Linguistics. 2018.

[29] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.