文章编号：1003-0077(2021)01-0034-09

基于大规模语料库的现代汉语动宾搭配知识库构建

王贵荣[2]，饶高琦[1,2]，荀恩东[2]

(1. 北京语言大学 汉语国际教育研究院，北京 100083；
2. 北京语言大学 信息科学学院，北京 100083)

摘 要：汉语缺乏词形变化，语法信息需通过词语搭配关系获得，且在生活中词语通常在搭配中发挥交际作用。因此无论是在语言学本体，还是在自然语言处理的各项任务中，词语搭配知识都尤为重要。各种搭配中，动宾搭配能够反映句子轮廓，并在数量和多样性方面具有优势地位，故该文聚焦于构建现代汉语动宾搭配知识库，以期为自然语言处理提供基础知识，同时也为语言本体研究、语言教学等提供大量实例。该文首先从语言本体的角度出发，总结了动宾搭配的知识体系，并根据该体系制定相应形式化检索式 140 个，从 BCC 语料库中抽取动宾搭配知识，并对抽取结果进行了初步消歧，最终获得动宾搭配 300 万对，形成动宾搭配知识库。

关键词：动宾搭配；知识抽取；知识库；BCC 语料库

中图分类号：TP391 **文献标识码**：A

Construction of Verb-object Knowledge Base from BCC Corpus

WANG Guirong[2], RAO Gaoqi[1,2], XUN Endong[2]

(1. Institute of International Chinese Language Education, Beijing Language and Culture University, Beijing 100083, China;
2. School of Information Science, Beijing Language and Culture University, Beijing 100083, China)

Abstract: The word collocation knowledge is essential to both linguistic ontology and natural language processing tasks, in which verb-object collocation is distinguished by its syntax role, its quantity and its diversity. This paper constructs a Chinese verb-object knowledge base to provide basic knowledge based on a large scale corpus. It first summarizes the knowledge system of verb-object collocation from the perspective of linguistic ontology, and formulates 140 queries to retrieve verb-object instances from the BCC corpus. Finally, three million pairs of verb-object collocation are obtained after disambiguation.

Keywords: verb-object collocation; knowledge extraction; knowledge base; BCC corpus

0 引言

在任何语言里，词语搭配都是一个重要问题，在汉语中，尤为突出。从语言本体来看，汉语重“意合”，词语缺乏形态变化，词与词的搭配有时会超出语法形式的约束，只要满足意义、逻辑的要求就可以搭配。因此，无论是语言本体还是语言教学研究，词语搭配都是言语组装的重要环节。从自然语言理解来看，要实现计算机对自然语言的“理解”，语言知识是必不可少的。而汉语缺乏形态变化，不能提供充分的形式化知识，因此，词语搭配知识也就成为自然语言理解各子任务的一个重要知识源。由于词语搭配描述的是词与词之间的组合情况，既包含结构知识也包含语义知识，更能准确地刻画出句子中词与词之间的联系，在句法分析中受到人们广泛的重视。依存句法认为，动词是句子的中心，依存分析中各个节点都是词，不存在词和短语或短语之间的关系判定问题，主要是通过获取句子的核心动词及其所支配的词语搭配，进而分析句子内词语之间的依存关系，以建立依存句法树。

一般而言，动宾结构在 SVO 型语言里是很常见的，是句内的核心成分，处于优势地位，可以形象地称之为“骨架”，它实际上映射了整个句子的轮廓。

收稿日期：2019-09-09 **定稿日期**：2019-10-19

基金项目：国家语委信息化专项项目(ZDI135-114)

1942年吕叔湘先生在《中国文法要略》中就指出“句子的中心是一个动词”[1]。1959年法国语言学家特思尼耶尔(L Tesniere)在“依存语法”的代表作《结构句法基础》中明确指出“动词是句子的中心，它支配着别的成分，而它本身却不受其他任何成分的支配。动词在句子中起的作用是关联，就是说动词把句子中其他的词连成了一个整体。”[2]只要能准确识别出动宾结构，就能在此基础上进行一些后续分析，从动词出发，可以向左识别各种状语，逼近句子的主语成分，从宾语出发，可以向左识别宾语的各种修饰成分，逼近动词，从而为实现深层句法分析奠定一定的研究基础。本文以大数据为支撑，构建动宾搭配知识库，以期为句法分析提供结构化引导知识，提高句法分析的准确率，同时该知识库也可为语言本体、语言教学研究等提供大量实例。

1　研究现状

1.1　语言学界的研究

一直以来，现代语言学界关于动宾搭配的研究就层出不穷，研究思路主要有四种。

一是在格语法的理论视角下，用宾语的语义角色的来对宾语进行分类，主要的研究有李临定[3]、马庆株[4]等；也有学者展开了宾语不同语义角色的细致研究，如宋玉柱[5]介绍了原因宾语的类型及不同类型中充当原因宾语的成分；陈昌来[6]否认了工具成分可以表现为主语、宾语，并介绍了工具成分可以出现的句法结构；赵旭[7]研究了处所宾语的判别标准、内部小类以及非典型处所宾语的生成动因。

二是以配价语法为理论支撑，从动词价位的角度来考察动词所带的宾语，如罗梦鹿[8]指出双宾语句式动词包括大部分三价动词和一部分二价动词；王慧[9]分析了二价动词不带宾语、带单宾语和带双宾语的情况；袁毓林[10]提出了一种基于配价层级和配位方式的汉语配价语法的描写模型，用以全面地反映动词在不同的句式中对名词性成分的支配能力及其句法组配方式。

三是从韵律的角度分析了动宾搭配的规律，如吕叔湘[11]指出汉语双音化倾向明显，并分析了单双音节对汉语划分词语边界的影响；冯胜利[12]系统阐释了韵律构词学和韵律句法学这两个全新的理论系统；骆健飞[13]指出单音节动词一般是强时空动词，倾向于搭配工具、方式类宾语，双音节动词一般是泛时空动词，倾向于搭配原因、目的类宾语。

四是从宾语的体谓性来考察动词特征，如宋玉柱[14]提出将动词按宾语的语法性质划分为体宾动词、谓宾动词和体谓宾动词三类；亢世勇[15]对常用谓宾动词带动宾、形宾、小句宾进行了分类统计；陈永莉[16]指出形式动词只能带双音节动词宾语，并介绍了形式动词受事成分的语法位置和宾语扩展形式；崔少娟[17]、孙萍[18]从动词分类、宾语语义特征等方面对《现代汉语动词用法词典》中的谓宾动词进行了全面研究；梁永红[19]研究了及物动词带名宾情况的发展变化的具体表现、特征以及影响因素。

从笔者的调研情况来看，目前已有的对动宾搭配的研究，基本都是选取动宾搭配的某一侧面进行定性研究，且文中也都是通过举例的方式来验证结论，尚未有人基于大数据对动宾搭配进行抽取和研究。

1.2　中文信息处理领域的研究

相对语言本体领域丰富多彩的研究，中文信息处理领域关于动宾搭配的研究则比较单一，主要是从动宾搭配的自动识别角度展开研究的，如孙宏林[20]从语料库中归纳了判断“V＋N”序列是合法短语的14条语法规则；高建忠[21]提出“匹配＋语义限制”和“匹配＋词语相似度”计算模型，用于动宾搭配的自动识别；李晋霞[22]从内部构成出发以定中“V_双＋N_双”结构类型的识别为突破口提出“V_双＋N_双”结构类型自动识别的规则；程月等人[23]提出机器学习中的条件随机场方法，用于汉语动宾搭配的自动识别。也有学者开始从语义的角度进行研究，如周卫华[24]从动宾之间的语义角色关系、动词对宾语的语义选择限制这两个方面详尽地考察了500个单音节动词和宾语之间的语义搭配情况；李斌[25]对动宾之间语义选择限制的多样性和强度差异做了系统标注和统计分析。

目前学者对动宾搭配所做的研究，无论是基于结构进行的对动宾搭配的自动识别和获取，还是跳过结构直接对动宾搭配进行语义分析和计算的，都是在探究动宾搭配的一种形式化规律，以方便计算机的处理，但是这种方法也只能覆盖语言中的一些高频现象。

此外，围绕中文信息处理构建的知识库也有很多。如由山西大学建设的汉语框架语义知识库(CFN)[26]是以加州大学伯克利分校的FrameNet为参照、以汉语真实语料为依据的供计算机使用的

汉语词汇语义知识库，主要包括框架库、句子库和词元库三部分。其中，词元库记录了词元的语义搭配模式和框架元素的句法实现方式。由北京大学开发的《现代汉语语法信息词典》[27]是为计算机实现汉语分析和汉语生成而研制的一部电子词典，全面地描述了所收录词语的语法信息。知网(HowNet)[28]是一个以汉语和英语的词语所代表的概念为描述对象，以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。北京大学袁毓林主持建设的《北京大学现代汉语实词句法语义功能信息词典》[29]，其中，《动词句法语义功能信息词典》主要包括动词的语法功能、语义角色及动词和语义角色组配的句法格式。句法组配格式提供句法上可搭配的位置，动宾搭配提供语义上可搭配的词语，二者配合使用，汉语句法语义分析将获得重大进步。也有学者编纂过搭配词典，如张寿康和林杏光主编的《现代汉语实词搭配词典》、赵培痒编著的《常用词语搭配词典》等。但它们只收录了高频使用的搭配，规模相对较小，不能全面反映动宾的搭配情况。目前已有的知识库主要提供语义、语法、常识知识，对动词各方面的知识有详尽描写，具有一定的形式化能力，一定程度上促进了中文信息处理技术的发展。但这些知识库仍以对语言学知识的抽象表征为主，计算机使用起来不够便捷。

本文拟基于BCC语料库构建动宾搭配知识库。BCC语料库语料来源领域较多，有文学、科技、报刊、博客等，能覆盖更多的语言现象，且BCC语料库处于动态更新状态，能及时捕获新的语言现象。基于该语料库构建的动宾搭配知识库更具全面性、时效性，对语言研究和语言教学而言，具有更高价值。动宾搭配知识抽取也是信息抽取的一项子任务，能够用于词义消歧、信息检索、机器翻译、句法分析、自然语言生成等多个方面。如词义消歧方面，人能够正确无误地理解某个词语，就是利用了词语的上下文信息，而词语搭配正是确定词语义项的上下文，动宾搭配知识为动词歧义消解提供了知识源。机器翻译方面，由于不同语言的搭配规律不同，造成不同语言词语之间的对等翻译极为困难，词语搭配的翻译将有助于提高翻译质量。句法分析方面，动宾搭配作为句法、语义双重关系的载体，能够帮助确定句子的核心成分关系，提高分析精度。

本文的研究将从以下几个方面具体开展：首先，从语言本体的角度出发，总结动宾搭配知识体系；其次，根据动宾搭配知识体系制定BCC语料库的检索式，获得动宾搭配知识对；最后，根据检索结果优化检索式，并对获得的动宾搭配知识进行消歧。

2 动宾搭配知识体系

前人从语言学角度对动宾搭配所做的研究相对较多，但前人研究多是从语义角度展开的，不利于形式化抽取。本文主要从宾语由哪些词类或结构充当的角度整理动宾搭配知识体系，首先从宏观的角度将动宾搭配分为三大类型，根据宾语的体谓性将动宾搭配分为“动＋体宾”和“动＋谓宾”两大类，又因为双宾语的情况比较特殊，故没有将“动＋双宾”归为上述两类中，单独归为一类；其次，对每一大类下的宾语构成做细致区分，并列举相应例句，该过程以朱德熙、李临定、刘月华先生的宾语语义分类为主，结合高校使用的现代汉语教材中对宾语构成成分的说明(表1)，最终形成一个动宾语义全面且便于形式化检索的知识体系。本文在抽取动宾搭配知识时，考虑到语言层层递归的复杂性以及抽取方式的局限，只对一般名宾、代词宾语、数量名宾语、简单定中作宾语、动词宾语、形容词宾语进行了抽取。

表1 前人的宾语分类体系

朱德熙	李临定	刘月华	黄廖本	北大本
受事宾语	受事宾语	受事宾语	名词	名词
施事宾语	对象宾语	结果宾语	代词	代词
工具宾语	处所宾语	工具宾语	数量词	偏正短语
结果宾语	结果宾语	处所宾语	的字短语	联合短语
终点宾语	工具宾语	方向宾语	联合短语	的字短语
时间宾语	目的宾语	原因宾语	动词	数量短语
处所宾语	原因宾语	目的宾语	形容词	动词(短语)
存现宾语	方式宾语	存现宾语	动宾短语	形容词(短语)
虚指宾语	致使宾语		主谓短语	主谓短语
程度宾语	角色宾语		复句宾主	
准宾语	动宾			
双宾语	小句宾			
谓词性宾语				

前人对体词性宾语的语义类研究较多，谓词性宾语一般分为动宾和小句宾。为便于书写形式化的检索式，本文从充当宾语的词类出发，对宾语进行重新分类，尽可能覆盖所有的宾语语义角色。宾语的常用语义角色基本是一般名词(n)作宾语，本文把

能用词性区分出来的处所宾语(ns)单独划分为一类，其余都归入一般名宾；数量结构作宾语，本文认为动量短语是对动作的补充说明，是补语而不是宾语，将名量短语和时量结构作宾语划分为数量宾语；代词和联合短语作宾语，根据体谓性将其分别归入体宾和谓宾中；存现句是一种特殊句法现象，且存现动词是一个封闭的类，故将存现宾语单独划分为一类。只有朱德熙先生的分类中有程度宾语，本文认为其是程度补语，不纳入宾语范围。体词性宾语中其他小类划分情况较为一致，这里沿用前人分类。谓词性宾语中，过去都只是粗略分为动宾、形宾、小句宾，本文将可以充当宾语的谓词性结构均单独分类，更加细致，便于从形式上区分。另只有黄廖本的《现代汉语》中提到复句作宾语的情况，鉴于复句也是谓词性的，将其划分为谓宾下的一类。双宾动词也是一个封闭词表，根据双宾中两个宾语的类型划分为“真宾＋准宾”“真宾＋真宾”两类。综上，本文定义的动宾搭配知识体系如表 2 所示。

表 2　动宾搭配知识体系表

动宾结构	宾语小类	举例
动＋体宾	一般名宾	“吃饭”“写信”
	代词宾语(名)	“相信他们”“你找谁?”“去那”
	数词宾语	“二加二等于四”
	数量名结构	“提高 M 个百分点”
	数量宾语	“住了三年”“等一会儿”
	处所宾语	“到家”“去南京”
	存现宾语	“黑板上写着字”“来了一个人”
	“的”字结构	“买木头的”“不吃冷的”
	定中结构	“穿新衣服”“得到群众的响应”
	联合结构(名)	“有哥哥、弟弟”
动＋谓宾	动词宾语	“喜欢笑”
	形容词宾语	“爱干净”“觉得很好”
	代词宾语(动)	“他同意这样吗”
	主谓结构	“希望他来”“听说她回来了”
	状中结构	“感到不舒服”“觉得十分快意”
动＋谓宾	动补结构	“同意研究一下”“感觉好点了吗”
	动宾结构	“看打球”“赞成讨论这个问题”
	联合结构	“给予表扬和鼓励”“他憎恶虚伪和怯懦”
动＋谓宾	连动结构	“他希望进来歇一会”
	兼语结构	“大家都主张邀请他参加会议”
	复句宾语	“发现风停了，浪也静了”
	v 一 v	“去看一看”
	v 了 v	“去看了看”
	vv	“去看看”
动＋双宾	真宾＋真宾	给予类动词　“送你一本书”
		获取类动词　“买了他一所房子”
		称谓类动词　“叫他老大哥”
	真宾＋准宾	“等他一会儿”“拍了他一下”

3　动宾搭配获取

3.1　检索系统

本文获取动宾搭配知识的语料库是对外开放的北京语言大学语料库 BCC(http://bcc.blcu.edu.cn)的延伸版，其包括报刊、文学、科技、微博等各领域的语料，数据规模较公开版更大，约 1.1 万亿字。该语料库能够支持集字符、属性和结构信息为一体的复杂查询，且检索速度较快。基于该语料库抽取动宾搭配知识，需要制定相应的 BCC 检索式，接下来将详细介绍检索式的构成。

3.1.1　简单检索

简单检索的检索式只有 Query 部分，可以包含字符串、词性符号、离合符号“＊”、单个词语标识符“～”、空格等内容，对上下文的限制较少，只能表达简单的结构和语义信息。BCC 简单检索式构成具体介绍如表 3 所示。

表 3　BCC 简单检索式构成说明

符号	含义	举例
字符串	中英文字符	“吃饭”，检索“吃饭”出现的情况
词性符号	主要包括 a、d、v、n、r、p、q、u、m、w 等，与北大词性标注集一致	“不 v”：检索“不”与动词共现的情况；“d v”：检索副词与动词共现的情况

续表

符号	含义	举例
*	一般情况下，检索式表示连续的语言片段，如果需要查找离合情况时，需要用“*”号。形式为“A*B”，表示查找 A 的后面离合出现 B 的单句	“洗*澡”：来检索“洗澡”离合出现的情况
～	表示一个词，在检索式中间数量不超过 5 个，在两边数量不限	“洗～澡”：表示“洗”+一个词+“澡”
空格	与通常搜索引擎含义不同，在模式中，有歧义表达时，起到分割作用。在检索式中，输入的英文字符与词性符号一致时，计算机处理为词性，否则按普通字符串处理，如果有表达歧义时，用空格分开	“一 q n”：检索“一”后面连着一个量，量词后面是一个名词的实例。多个词性相连时，用空格隔离

3.1.2 复杂检索

复杂检索的检索式包括 Query 和 Condition 两部分，形如“Query{condition1；condition2；...；print($i)}”。从功能上看，复杂检索式可以对上下文进行条件约束，对抽取部分进行韵律结构限制、词属性类限制，同时可以实现同一个检索式中词表的实例化检索，提高检索效率。从形式上看，复杂检索式的 Query 部分可以出现“()”，表示被限定的部分，condition部分表示条件限制，print 表示输出语句。如“不(v)(n) W{len($1)=2；len($2)=2；print($1$2)}”，表示“不+双音节 v+双音节 n+标点”共现的情况，用“()”括起来的部分表示需要予以限定的部分，“$1”表示第一个被括起来的部分，可以用词表对其进行类的限制，“[S_T_体谓准_体]”为自定义词表，表示体宾动词，也可对音节进行限制，“len($1)=2”表示第一个元素即“v”是双音节的；“W”表示标点符号，这里是指以标点结尾的动名搭配；“print($1$2)”这里表示输出 query 部分被括起来的内容，即只输出“v n”搭配，没有 print 语句时，默认输出整个检索式的检索结果。同时，复杂检索式可以使用“$V”表示实例化检索式词表中的词，如“$V=[S_V_趋_趋向动词]”，表示将趋向动词表中的词语逐个放入检索式中“$V”的位置进行检索。

3.1.3 简单检索与复杂检索抽取结果对比

简单检索式抽取结果和复杂检索式抽取结果对比如表 4 所示。

表 4 简单检索与复杂检索抽取结果对比

	检索式	检索例举	频次
简单检索	v n	逛街	730 471
		看电影	695 285
		*采访时	643 536
		分享图片	629 963
		*参与方式	542 207
	打*n	打*电话	15 030
		打*球	2 795
		打*热线	1 912
		打*人	1 886
		打*比赛	1 591
复杂检索	不(v)(n) W{len($1)=2；len($2)=2；print($1$2)}	需要理由	33 901
		喜欢世俗	26 433
		相信爱情	24 459
		知道真假	10 665
		相信眼泪	10 057
	$V过(n) W{$V=[S_V_趋向动词_趋]；len($1)=2}	上过大学	4 333
		出过家门	1 569
		出过问题	1 388
		去过医院	994
		下过大雨	648

注：“*”在文字前表示该示例不是动宾结构；“*”在文字之间表示检索式中的离合符号。

简单检索式“v n”抽取的搭配中“采访时、参与方式”并不是动宾搭配，“采访时”的韵律构成是 2+1，冯胜利[12]指出“2+1”式动宾组合容易导致“抑扬”结构，普通重音无法实现，不易构成动宾搭配，因此，可以分别采用不同的韵律构成单独检索。“参与方式”是动词作定语构成的定中结构，而动词作定语时，一般是不能被否定副词修饰的，可以在检索式中加入否定副词“不”进行限制，如复杂检索式“不(v)(n) W{len($1)=2；len($2)=2；print($1$2)}”，则可以避免上述情况的出现，提高检索的准确性。简单检索式“打 * n”只能抽取关于动词“打”的离合型名词宾语，而不同属性类的动词带宾的上下文情况不同，需要根据动词小类及宾语上下

文情况细化检索式。复杂检索式“$V 过(n) W {$V=[S_V_趋向动词_趋];len($1)=2}”中，“$V”表示动词某一属性类的词表，该检索式可以对词表内的词语进行实例化检索，大大提高检索效率。经过对比可知，复杂检索式效果要好于简单检索式，故本文在抽取时大多采用复杂检索式。

3.2　动宾搭配规则

上一节详细介绍了动宾搭配的分类情况，本节将详细说明为抽取动宾搭配知识制定的检索式情况。现代汉语语法具有递归性，不同结构类型层层嵌套形成的动宾结构比较复杂，故本文目前只抽取了简单类型的宾语，即体词性宾语中的一般名宾、代词宾语、数量名宾语的连续类型和离合类型，谓词性宾语中的动词宾语和形容词宾语的连续类型和离合类型。根据这几种情况，再分别从属性类、上下文、韵律结构和自然标注信息等方面添加限制条件，共制定检索式223个。

3.2.1　连续型动宾搭配规则

连续型动宾主要抽取了动词后紧邻宾语的情况，在检索时重点关注动词的上下文，从动词的修饰语、属性类、动宾的韵律构成和标点信息等方面来添加限制条件，尽可能使抽取的搭配能够构成动宾关系。连续型动宾检索情况如表5所示。

表5　连续型动宾检索情况

类型	检索内容	举例
用标点W限定，避免宾语位置的词语与后一词语构成搭配	以标点结尾的v n	“打击敌人”
	以标点结尾的v r	“相信自己”
	以标点结尾的v v	“准备睡觉”
用音节限定	双音节v+双音节n	“解决问题”
	单音节v+单音节n	“买菜”
	单音节v+双音节n	“接电话”
	双音节v+双音节v	“感谢分享”
	单音节v+双音节v	“爱打架”
	双音节v+双音节a	“感觉不错”
用词属性表限定	体宾动词+n	“看电影”
	体宾动词+r	“打他”“吃什么”
	系动词+m	“等于四”
	谓宾动词+v	“禁止停车”
	谓宾动词+a	“给予幸福”
用“不、没、很”限定	不+体宾动词+n	“不知道真相”“不打架”
	没+体宾动词+n	“没发现问题”“没吃饭”
	很+心理动词 +n	“很喜欢白色”
	很+心理动词 +v	“很喜欢打球”

3.2.2　离合型动宾搭配规则

离合型动宾主要抽取了动词和宾语之间有其他词语出现的情况，抽取时重点关注能出现在动宾之间的不同离合成分，抽取了离合成分为“着了过”、“了个”、数量、宾语的定语成分等的动宾搭配。离合型动宾检索情况如表6所示。

表6　离合型动宾检索情况

类型	检索内容	举例
“着了过”离合	体宾动词+着+n	“听着音乐”
	体宾动词+了+n	“打了电话”
	体宾动词+过+n	“参加过战争”
	谓宾动词+了+v	“避免了挥霍”
	谓宾动词+了+v	“感到了恐惧”
	谓宾动词+着+v	“象征着勇敢”
“了个”离合	双音节v+了个+n	“发现了个秘密”
	单音节v+了个+n	“做了个梦”
	v+了个+一个词+的+n	“取了个好听的名字”
“m、q、mq”离合	动词+量词+名词	“请教个问题”
	动词+数词+名词	“持续三小时”
	动词+数词+量词+名词	“买两斤大米”
“～的”离合	动词+n的n	“远离城市的喧嚣”
	动词+r的n	“牵你的手”
	动词+一个词+的n	“喜欢漂亮的衣服”

3.3　动宾搭配消歧

在第一轮抽取工作结束后，笔者详细观察了抽取到的动宾搭配知识，发现抽取的知识长尾效应明显，且由于语料的分词错误、词性标注错误和检索式的局限性，抽取到的动宾搭配数据中也存在着一些非动宾搭配的类型。动宾搭配知识作为句法分析中最基础的资源，其准确性直接影响整个句法分析器

的效果，因此，为了获得更为准确的动宾搭配知识，本文从检索式书写、动词、宾语等方面进行了初步的消歧。

3.3.1 检索式优化

为提高检索结果的准确率，笔者对初步制定的223个检索式人工进行了有效性评估，分别用1到5来表示检索式有效性从低到高，对于有效性低于3的检索式从限制动词和宾语两个方面进行改进，若改进后检索效果有所提升，则保留改进后的检索式，若改进后检索效果仍不理想，则舍弃该检索式。如简单检索式“(v)(n) W{len($1)=2;len($2)=2}”的有效性只有2，虽然该检索式能够召回大量的“VN”对，但非动宾搭配的负例情况也较多，比如“联系电话”“购买地址”这种最典型的动词作定语修饰名词的例子也会被当作动宾搭配抽取出来，故在动词前用典型否定副词“不、没”加以约束，并对“V”和“N”进行属性类的约束，构造出更有效的检索式“不(v)(n)W{$1=[S_V_体谓准_体];$1!=[S_V_趋向动词_趋];len($1)=2;$2=[P_N_宾语_可];len($2)=2;print($1 $2)}、没(v)(n)W{begin($1)!=[有];$1=[S_V_体谓准_体];len($1)=2;$2=[P_N_宾语_可];len($2)=2;print($1 $2)}”，一定程度上减少了非动宾搭配对。经评估改进后，共得到140个检索效果相对较好的检索式。

3.3.2 动词部分消歧

动词部分引起歧义主要是由两方面的原因导致，一是动词方面，即动词不能带宾语或抽取出来的是动词作定语的情况；二是语料库方面，即BCC语料库的分词错误、词性标注错误及分词粒度等原因。

针对动词方面的原因，笔者在抽取语料时根据前人研究整理了及物动词表、体宾动词表、谓宾动词表、可作定语的动词表、《现代汉语词典(第7版)》中的动词表、心理动词表、趋向动词表等一系列动词子类表。一方面，在书写检索式时可以使用这些词表作为限制条件，提高检索效果；另一方面，可以对抽取结果进行筛选。如“不起精神”虽然符合检索式“不(v)(n)W{$1=[S_V_体谓准_体];len($1)=1;$2=[P_N_宾语_可];len($2)=2;print($1 $2)}”，“起”也可以带体宾，例如，“起作用”“起血泡”等，但观察语料发现，“不起精神”并不是动宾搭配，而是“打不起精神”的一部分，而且“起”作补语的情况要更为普遍，所以笔者利用趋向动词表将趋向动词的搭配从检索结果中抽取出来，人工校验。

针对语料库方面的原因，笔者以《现代汉语词典(第7版)》的动词为标准，将与词典词性不一致的视为词性标注错误，但也有一些特殊情况除外。如词典中没有“看到”一词，这主要是因为“看到”可以理解为动词“看”与趋向动词“到”组合形成的述补结构，但由于二者结合比较紧密，高频使用，故语料库往往将其切分为一个词。针对这种不一致，仍保留该词为动词。而“达”在词典中为一个语素，但是在语言中经常会有“人口达13亿”“产值达290亿元”“竹制品已达200多个”等“达”作动词，后常跟数量短语的用法，因此也将其视为一个词。“把把”应该是“把把关”，是“把关”一词的变形，虽然语料库中将“把把”切分为一个动词，但抽取动宾搭配时不宜将其视为一个词。

3.3.3 宾语部分消歧

宾语部分引起歧义也分为宾语自身和语料库两方面的原因。前者主要是宾语部分不能与动词构成动宾搭配，如“时候”“台风”“产品”不能与动词“打”构成动宾搭配，但这类现象几乎在每个动词的搭配表中都会出现，分布较为离散，本文目前只将低频部分舍去，尚未对高频部分进行过滤。后者主要也是分词错误和词性标注错误。经观察语料发现词性标注错误主要表现为英文字母、标点符号、数字、其他词性的词等都有被标为名词的现象，比如语气词“吗”、代词“那”等。分词错误主要表现为把标点和词语切分在一个词语内，如“W酸奶”“眼病W”等。对于词性错误和标点切分错误，统一采用正则表达式对抽取结果进行剔除。

3.3.4 人工校对

正如齐夫律(Zipf's Law)揭示的那样，针对于一种语言的词汇分布来说，极少数高频词(型)的出现次数已经覆盖一个语料库总词数的绝大部分，而词(型)总数中大约一半的词(型)在这个语料库中却只出现一次。词语搭配的分布同样也遵循齐夫律，因此，本文在上述消歧结束后选取了动宾搭配中高频80%的部分，进行了人工消歧，最终获得动宾搭配300万对。动宾搭配知识库各子类分布情况如表7所示。

表 7　动宾搭配知识库各子类分布情况

动宾搭配	搭配对数	动词数	宾语类型	检索式个数	语料条数	占比/%
动＋体宾	5 760 639 (89.74%)	5 856	连续型名宾	80	42 977 937	0.94
			离合型名宾	28	1 362 558	0.03
			代词宾语(体)	4	10 774	0.001
			简单定中作宾语	4	1 045 309	0.023
			数量名宾语	4	281 631	0.006
动＋谓宾	658 581 (10.26%)	829	连续型动词宾语	9	2 053 745	0.876
			连续型形容词宾语	5	247 757	0.106
			离合型动词宾语	2	11 472	0.005
			离合型形容词宾语	4	29 650	0.013

从表 7 可知，能够带体词性宾语的动词数量要比能够带谓词性宾语的动词多，动宾搭配知识库中“动＋体宾”的搭配对数占总搭配数的 89.74%，要远远高于“动＋谓宾”的 10.26%，这说明了体词比谓词更容易被支配，人们在语言生活中表达较多的是动作行为与客观事物、对象的关系，以及人们对客观事物、对象的观点、看法等；表达较少的是动作行为与动作行为的支配关系。其中，体词性宾语中连续型名宾的数量最多，占了体宾总数的 94%；其次是离合型名宾，占体宾总数的 3%，如图 1 所示。谓词性宾语中连续型动词宾语的数量最多，占了谓宾总数的 87.6%，其次是连续型形容词宾语，占谓宾总数的 10.6%，如图 2 所示。体宾与谓宾相比，离合型宾语更多，即“动＋体宾”中更容易添加“着、了、过”等词语，以表示动作发生的时态，而“动＋谓宾”中，动词大多数是心理动词，时态性较弱，更倾向于紧邻搭配。

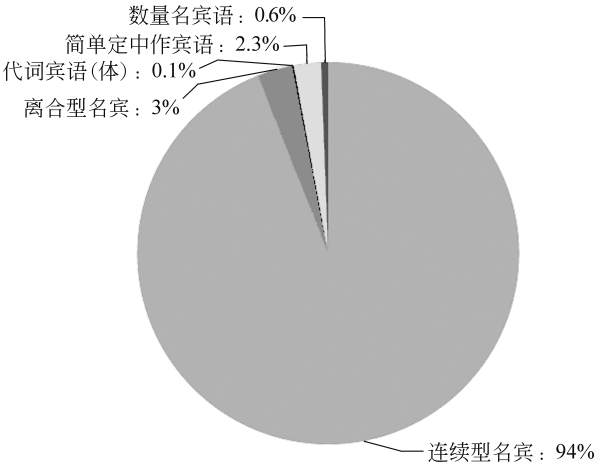

图 1　体宾各子类分布情况

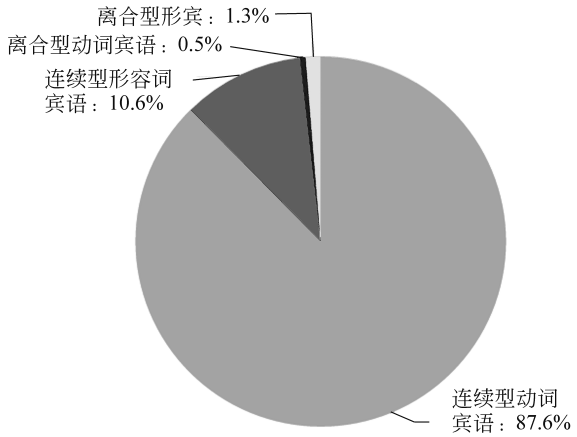

图 2　谓宾各子类分布情况

3.4　动宾搭配知识库

本文将抽取到的 300 万对动宾搭配按层级入库，即先按“动＋体宾”“动＋谓宾”分类，再将每一类下每个动词的所有宾语按频次高低排序，具体如图 3

.VN_丢失	
(	
[	
钱包	8225
地点	6954
爱情	5027
事件	4188
自己	3529
物品	2915
东西	2852
时间	2527
梦想	2386
自行车	2143
数据	2064
水分	1902
自我	1780
细节	1707
阵地	1572
护照	1517
钥匙	1500
问题	1420

.VP_下令	
(	
[	
撤退	6497
开枪	3378
开车	3009
逮捕	2634
禁止	2131
拆除	1831
开火	1806
停办	1487
处死	1481
删除	1405
说	1273
退兵	1053
解散	1022
停止	1021
疏散	996
拦截	975
回收	966

图 3　动宾搭配库存储形式

所示。“VN”表示“动＋体宾”,“VP”表示“动＋谓宾”。且本文抽取的动宾搭配已建立检索服务,可以查询某一个动词的全部宾语,也可以查询两个词语是否是动宾搭配。

4 结语

本文主要从动宾搭配知识体系的构建、检索式构成及书写、动宾搭配抽取及消歧等方面展开工作,共制定检索式 140 个,抽取到动宾搭配 300 万对,构建了一个规模较大、质量较高的动宾搭配知识库。该知识库的构建不仅为中文信息处理的子任务提供了大规模、高质量的基础知识,提高了计算机分析语言的能力,同时也为语言研究和语言教学提供了大量真实可靠的实例。此外,在构建动宾搭配知识库的过程中发现,尽管动宾搭配的知识体系较为完善,但汉语缺乏形态变化,很多语言学知识人们能够理解并很好地运用,但却无法将其形式化,转化为计算机可利用的知识。因此,本文的工作仍有一定的不足之处。首先,本文利用更多的是词性信息、动宾搭配的韵律条件及少量的动词子类信息,只完成了简单动宾搭配的抽取,对层层嵌套递归性的动宾抽取尚无能为力。其次,检索式自身的表达能力也相对有限,在抽取动宾搭配知识时,只能体现有限的上下文,且语料库自身存在着分词和词性标注的错误,造成后期消歧压力较大。最后,由于人力物力的限制,本文只对抽取结果进行了初步消歧,检索结果仍有进一步消歧的需要。

目前,本文初步完成了动宾搭配知识库的构建,今后还可以从以下几个方面进一步完善和改进。第一,采用计算的方法对抽取结果再次进行消歧,提高动宾搭配知识库的质量;第二,利用已有知识库建立深度学习模型,自动抽取本文目前尚未覆盖的其他动宾搭配类型,不断完善动宾搭配知识库;第三,探索将动宾搭配方面更多语言知识形式化的方法,降低知识抽取的难度。

本文资源将逐步以合宜方式在学术界和工业界共享。

参考文献

[1] 吕叔湘.中国文法要略[M].北京:商务印书馆,1942.

[2] 宋丽.特斯尼耶尔《结构句法基础》的要点梳理及简评[J].文教资料,2017(Z1):30-34.

[3] 李临定.宾语使用情况考察[J],语文研究,1983(02):31-38.

[4] 马庆株.名词性宾语的类别[J].汉语学习,1987(2):3-8.

[5] 宋玉柱.略谈原因宾语[J].南开学报,1980(5):48-50.

[6] 陈昌来.工具主语和工具宾语异议[J].世界汉语教学,2001(1):65-73.

[7] 赵旭. 现代汉语处所宾语研究[D].杭州:浙江大学硕士学位论文,2010.

[8] 罗梦鹿. 双宾语句式动词的配价研究[D].成都:四川师范大学硕士学位论文,2007.

[9] 王慧. 二价动词与宾语[D].沈阳:辽宁师范大学硕士学位论文,2007.

[10] 袁毓林.汉语配价语法研究[M].北京:商务印书馆,2010.

[11] 吕叔湘.现代汉语单双音节问题初探[J].中国语文,1963(01):10-22.

[12] 冯胜利.汉语的韵律、词法与句法[M].北京:北京大学出版社,2009.

[13] 骆健飞.论单双音节动词带宾的句法差异及其语体特征[J],语言教学与研究,2017(1):14-24.

[14] 宋玉柱.关于体宾动词和谓宾动词[J],世界汉语教学,1991(2):90-91.

[15] 亢世勇.现代汉语谓宾动词分类统计研究[J].辽宁师范大学学报,1998(01):36-39.

[16] 陈永莉.形式动词后带宾语的多角度研究[J].安徽教育学院学报,2006(02):93-95.

[17] 崔少娟. 现代汉语谓宾动词研究[D].广州:暨南大学硕士学位论文,2011.

[18] 孙萍. 现代汉语谓宾动词研究[D].沈阳:辽宁大学硕士学位论文,2015.

[19] 梁永红.当代汉语及物动词带名宾情况的发展变化:基于《动词用法词典》的定量统计分析[J].语文研究,2017(04):20-26.

[20] 孙宏林.从标注语料库中归纳语法规则:"V＋N"序列实验分析[C]//语言工程.全国第四届计算语言学联合学术会议论文集,北京:1997:157-163.

[21] 高建忠.汉语动宾搭配的自动识别研究[D].北京:北京语言大学硕士学位论文,2000.

[22] 李晋霞. 面向计算机的“V_双＋N_双”结构类型研究[A]. 语言文字应用研究论文集(Ⅱ)[C].教育部语言文字应用研究所,2004:7.

[23] 程月,陈小荷. 基于条件随机场的汉语动宾搭配自动识别[J]. 中文信息学报, 2009, 23(1): 9-15.

[24] 周卫华. 面向中文信息处理的现代汉语动宾语义搭配研究[D].武汉:华中师范大学博士学位论文,2007.

[25] 李斌.动宾搭配的语义分析和计算[M].北京:世界图书出版公司,2011.

[26] 郝晓燕,刘伟,李茹,等.汉语框架语义知识库及软件描述体系[J].中文信息学报,2007(05):96-100,138.

[27] 俞士汶,朱学锋,王惠,等.现代汉语语法信息词典规格说明书[J].中文信息学报,1996,06(02):1-22.

[28] 董振东,董强.知网和汉语研究[J].当代语言学,2001(01):33-44,77.

(下转第 53 页)

matching energy function for learning with multi-relational data[J]. Machine Learning, 2014, 94(2): 233-259.
[28] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data[C]// Proceedings of the International Conference on Machine Learning, 2011: 809-816.
[29] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases[C]//Proceedings of AAAI, 2011: 301-306.
[30] Wu J, Xie R, Liu Z, et al. Knowledge representation via joint learning of sequential text and knowledge graphs[J]. arXiv preprint arXiv:1609.07075, 2016.
[31] An B, Chen B, Han X, et al. Accurate text-enhanced knowledge graph representation learning[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018.


翟社平(1971—)，博士，副教授，主要研究领域为自然语言处理和机器学习。
E-mail: s2422437140@tom.com


王书桓(1996—)，硕士研究生，主要研究领域为知识图谱和表示学习。
E-mail: 17835422718@163.com


尚定蓉(1994—)，硕士研究生，主要研究领域为自然语言处理。
E-mail: sdr_person@163.com

(上接第 42 页)

[29] 李强，袁毓林.生成词库理论和名词语义的结构描述与概念解释[C]//词汇学国际学术会议暨第十一届全国汉语词汇学学术研讨会，北京：北京大学，2016
[30] 林建方.词搭配抽取及在信息检索中的应用研究[D].哈尔滨：哈尔滨工业大学博士学位论文，2010.
[31] 朱德熙.语法讲义[M]，北京：商务印书馆，1982.
[32] 刘月华，等.实用现代汉语语法[M]，北京：商务印书馆，2001.
[33] 李临定.动词的宾语和结构的宾语[J]，语言教学与研究，1984(3)：103-114.
[34] 孟琮.动词用法词典[M].上海：上海辞书出版社，1987.
[35] 崔玲齐. 类型学视角下的现代汉语谓宾动词研究[D].上海：上海师范大学博士学位论文，2012.
[36] 荀恩东，饶高琦，肖晓悦，等.大数据背景下 BCC 语料库的研制[J]，语料库语言学，2016(1)：93-109.
[37] 李晋霞. 现代汉语定中“V 双＋N 双”结构研究[D].北京：中国社会科学院研究生院博士学位论文，2002.
[38] 张寿康，林杏光. 汉语实词搭配词典[M].北京：商务印书馆，1992.
[39] 俞士汶，等.现代汉语语法信息词典详解[M].北京：清华大学出版社，1998.
[40] 中国社会科学院语言研究所词典编辑室.现代汉语词典(第 7 版)[M].北京：商务印书馆，2016.
[41] 黄伯荣，廖序东.现代汉语[M].北京：高等教育出版社，2017.
[42] Lucien Tesniere. Elements of Structural Syntax[M]. John Benjamins Publishing Company, 2015.


王贵荣(1994—)，硕士研究生，主要研究领域为计算语言学、句法语义分析、语言资源建设、搭配库构建。
E-mail: guirongwang@126.com


饶高琦(1987—)，博士，助理研究员，主要研究领域为计算语言学、语言规划学、数字人文。
E-mail: raogaoqi@blcu.edu.cn


荀恩东(1967—)，通信作者，博士，教授，主要研究领域为自然语言处理、基于汉语大数据语言知识抽取、汉语句法语义分析、语言资源建设。
E-mail: edxun@blcu.edu.cn