文章编号：1003-0077(2020)11-0001-08

# 基于大规模语料库的介词结构搭配库构建

邢 丹[1]，饶高琦[1,2]，荀恩东[1]，王诚文[1]

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

**摘 要**：语言知识可帮助计算机正确地处理自然语言，介词结构知识作为语言知识的一种，对自然语言处理和语言教学研究有很重要的意义。该文基于大规模语料库构建了高质量的介词结构搭配库。首先在前人研究的基础上，对介词进行归类并建立了介词搭配知识体系，而后设计并实现了从大数据中获取介词结构搭配知识的规则，最后对抽取结果及其数据规模进行了统计和评估，主要目的是通过形式手段获取高质量的介词结构搭配，同时也为自然语言处理和语言学基础研究及应用提供数据支持。

**关键词**：语料库；知识抽取；介词结构搭配

**中图分类号**：TP391 **文献标识码**：A

## Large-scale Corpus Based Preposition Structure Collocation Base

XING Dan[1], RAO Gaoqi[1,2], XUN Endong[1], WANG Chengwen[1]

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

**Abstract**: Prescription structure is of great significance to natural language processing and language teaching research. This paper constructs a high-quality preposition structure collocation base from large-scale corpus. First, we determine the classification scheme of preposition and collect preposition collocation. Then, we design and acquire rules of prepositional structure collocation from large data. Finally, we test and analyze the extracted result.

**Keywords**: corpus; knowledge extraction; preposition structure collocation

## 0 前言

搭配是两个词之间的组合，其介于词和短语之间，在词和短语之间架起了一个桥梁。根据搭配组成部分之间的句法结构关系，可以将其分为主谓搭配、动宾搭配、定中搭配、状中搭配等类型[1]，而介词短语作为状中搭配的一种，由于内部构成相当复杂、兼类介词存在、介词短语本身存在歧义等原因[2]，导致介词短语一直以来都是自然语言处理的难点之一。搭配知识库的建设可以大幅提高计算机处理语言的能力，同时也可以为语言本体、教学及应用研究提供搭配案例。本文介词结构搭配库的资源构建工作分成三部分：首先，在前人研究的基础上构建了介词结构搭配知识体系；其次，设计并实现了从大数据中获取介词结构搭配知识的规则；最后对抽取的搭配知识进行统计和评估。

## 1 研究现状

搭配一词最早由语言学家 Firth 提出，他把搭配看成是词汇层面的一种语言现象，强调词汇间的共现关系，并把搭配整体看成表达意义的方式[3]。但是 Firth 并没有给出明确的定义，Halliday 在词语学框架下给出了搭配的定义，许多学者在这一定义的基础上继承并发展了搭配的定义[4-7]，Xu 等人

收稿日期：2019-09-01 定稿日期：2019-10-09

基金项目：国家重点研发计划“云计算和大数据”重点专项项目(2018YFB1005105)

给出了更加具体的定义,“搭配是包含两个或多个具有句法或语义关系的词的组合”[8]。国内对于搭配问题的研究,早期主要集中于探讨词语搭配的本质是语法关系还是语义关系[9-12],由于缺乏定量分析,目前还未形成普遍共识[13]。此外,还出现了一批实践性的成果。例如,张寿康、林杏光编纂的《学生常用词语搭配词典》《简明汉语搭配词典》《现代汉语实词搭配词典》等[14-16]。

以往对搭配知识抽取的研究,主要是基于词共现和分布值[17-18],也有基于规则的。例如,Kilgarriff等人通过词性标签构建了搭配规则建立了Word Sketch Engine系统[19];Huang等人将Sketch Engine扩展成中文,发现基于词性规则可以有效地提取语法信息[20];Hu和Smadja等人基于句法分析的方法从语料库中抽取搭配[21-22]。对于介词短语获取的研究也主要分为基于规则的方法和基于统计的方法。李洪政等总结了目前汉语介词短语识别主要存在的问题,并详细梳理了近些年介词短语研究的方法,其中包括规则方法、统计方法,再到无监督和有监督学习的方法等[23]。例如,规则方法有郑州大学自然语言处理实验室构建介词用法规则库来进行介词结构的识别,北京师范大学中文信息处理研究所构建的汉语专利语料知识库中总结了不同配价动词可以与哪类介词一起使用的搭配特征[24]。基于统计的方法主要使用机器学习模型进行自动识别。例如,温苗苗等人、鉴萍等人、卢朝华等人使用SVM建立了介词结构的自动识别系统[24-26];干俊伟和黄德根、奚建清和罗强、Li等人利用HMM模型进行介词短语识别[27-29];霍亚格和黄光君等人使用最大熵模型进行介词短语识别[30];近年来涌现了使用CRF模型进行介词短语识别的研究[30-35]等。

本文基于BCC语料库[35],根据语言学规律构造了形式化的知识检索规则,开展了介词结构搭配知识抽取工程。由于BCC是不定期更新语料的,语料相对较新,抽取出的数据相对于过去的研究具有相对的时效性,同时在利用词性标签的基础上,利用语言学总结的词长、停顿、韵律结构等规律特征构造检索规则,方法上有所创新,为进一步基于统计和机器学习的研究提供了大规模真实数据,也为语言本体、教学及应用研究提供了搭配实例,从而对自然语言处理基础和应用领域及语言学研究具有实际意义。

## 2 介词结构搭配知识体系的构建

本文的研究对象是由介词引导动词论元、以动词为中心语的介词结构搭配,例如,“跟老师聊天”“从北京出发”“用大碗吃饭”等这样的无递归、不嵌套的简单介词结构搭配。对于介词结构,前人已有一些研究。例如,“现代汉语虚词讲义”根据介词的语法功能介引对象的不同,将介词分为时间、处所、方向、对象、凭借依据、原因目的六类[37]。马杜鹃和郑通涛通过5本大纲对比,选出了60个介词,包括15个时空类介词,16个对象类,13个依据类,4个缘由类,8个施受类,4个其他类介词[38]。傅雨贤将介词分为前置词和后置词,他将前置词分为施事、受事、工具、对象内容、时空、方式依据、排除介词、原因目的、比况[39]。陈昌来提出类似的概念,称之为介词框架,认为介词是介词框架的前部,与介词相搭配的是介词框架的后部,他将后部词语的情况分为方位词、名词短语、连词、动词、介词、准助词[40]。本文在此基础上,构建了自己的介词分类体系,根据介词引导动词论元角色的不同将其分为13大类22小类,共总结了134个介词。介词的归类都是参考相关学者的研究总结归纳的[36-39],如表1所示。

表1 介词归类表

| 论元角色 | | 介词 |
|---|---|---|
| 主体 | 施事、系事 | 被 叫 让 给 由 归 挨 捱 叫 任 一任 任凭 任着 听 听任 听凭 随 于 为 |
| 客体 | 受事、成事、感事 | 把 将 对 对于 |
| 邻体 | 与事、当事、共事 | 与 替 为 同 给 和 跟 给 管 向 对 对于 就 论 拿 经过 就 关于 对着 管 从 冲 冲着 代 当 当着 对着 把 朝 朝着 连 连同 面对 随 随同 随着 引 以 针对 针对着 |
| | 同事 | 除 除了 除开 除去 |
| | 基准 | 比 较 较之 比较 比起 比起…来 比之 比之于 跟 和 较 较之 同 与 像 |
| 用事 | 工具、材料、方式 | 用 拿 以 把 将 靠 靠着 挨 挨着 拿 通过 经过 由 |
| 方面 | | 按 按着 按照 从 对于 关于 就 就说 论 拿 围绕 围绕着 在 照 照着 至于 作为 关于 对于 至于 拿 就 就说 论 |
| 条件 | | 趁 趁着 乘 即 借 借着 经 经过 经由 就 就着 冒 冒着 顺 俟 随 随着 在 |

续表

| 论元角色 | | 介词 |
|---|---|---|
| 依据 | | 按 按照 依 依照 照 据 依据 根据 以 凭 由 拿 趁 凭着 本着 通过 经过 随着 从 遵照 鉴于 按着 本 冲 冲着 基于 靠着 如 顺着 依仗 依仗着 依着 仗着 照着 论 随 因 |
| 源点 | | 从 自 由 打 自打 自从 打从 于 起 |
| 终点 | | 在 到 向 往 即 至 |
| 目的 | | 为了 为 为着 |
| 原因 | | 由于 由 以 因 因为 |
| 处所 | | 自 在 于 由 打 离 距 向 距离 离 往 到 奔 朝 冲 对 沿 沿着 顺 顺着 |
| 时间 | | 自 在 于 由 打 离 距　怕 向 自从 从 离 距离 |

本文中的介词结构搭配主要以动词为中心，通过明确介词分类体系，根据介词介引对象的不同，将介词实例化与动词搭配组合构造检索式，建立了介词与动词的搭配抽取体系，如表 2 所示。

表 2　介词与动词的结构搭配抽取体系

| 介词结构搭配体系 | 释例 |
|---|---|
| 实例化介词＋名词＋动词 | 跟老师报告 |
| 实例化介词＋名词＋动词＋标点 | 用粉笔画画 W |
| 实例化介词＋代词＋动词 | 和我聊天 |
| 实例化介词＋任意词(～)＋实例化方位词＋动词 | 朝河对面走 |
| 实例化介词＋任意词(～)＋实例化方位词＋动词＋标点 | 到清朝末期结束 W |
| 实例化介词＋处所词＋动词 | 从北京飞 |
| 实例化介词＋处所词＋动词＋标点 | 从北京出发 W |
| 实例化介词＋时间词＋动词 | 在四月开始 |
| 实例化介词＋时间词＋动词＋标点 | 于十二点出发 W |

观察表 1，我们发现同一介词可能引导不同论元角色，例如，“于”既引导源点又引导时间、处所等。这在一定程度上造成了歧义，给后期处理带来困难，基于这种现状，本文在北大“现代汉语信息词典”基础上[41]，对介词进行了再细分，如表 3 所示。“P_处所_指代处所_处_v”指能介引处所词后也能带动词的介词表，“P_地点_s_v”指的是能介引地点词后能跟动词的介词表，“P_人称代词_v_pr”指的是能介引人称代词后能跟动词的介词表，“P_时间_v_ptv”指能介引时间词后跟动词的介词表。

表 3　部分介词再细分归类表

| 类别 | 介词 |
|---|---|
| #P_处所_指代处所_处_v | 把 朝 朝着 从 打 当 到 顺 顺着 往 向 向着 沿 沿着 由 在 |
| #P_地点_s_v | 朝 朝着 顺着 往 向 向着 沿 沿着 由 在 |
| #P_人称代词_v_prv | 把 被 朝 除 除了 给 跟 管 归 和 将 叫 靠 冒 拿 凭 凭着 让 替 同 为 为了 向 以 用 由 于 在 照 照着 |
| #P_时间_v_ptv | 趁 从　打 待 到 到 于 在 至 自 自从 |

## 3　介词结构搭配知识的获取

### 3.1　检索式的构成

介词结构搭配获取是基于北京语言大学 BCC 语料库进行的，BCC 包含语料 150 亿字，包括报刊、文学、微博、科技、综合、古代汉语等多领域语料[35]，虽然介宾结构在不同语体中的表现不同，但是本文暂不考虑语体的不平衡性，以期构造一个通用的介宾结构搭配库。为从搭配库抽取高质量的知识，本文基于语言规则使用了一系列规则化的检索式。基本检索式构成如表 4 所示。

表 4　检索式构成

| 类型 | 描述 |
|---|---|
| 中英文字符 | 长度是否有限制 |
| 词性符号 | 沿用 BCC 语料库的词性表 |
| 空格 | 字母与字符串连用时，字母将不会被当作词性检索符号，空格用于词性符号之间或词性符号与汉字串之间的隔离 |
| * | 表示离合结构，且离合符号前后占位符不超过 5 个，离合结构内部字符数不超过 5 个 |
| ～ | 表示一个词，检索式中可包含多个 |

检索式在基本检索式的基础上增加了条件语句或输出语句。语句之间用“;”隔开，写在检索式后的“{ }”中，形如 Query{cond1;cond2;...;cond$i$;print($i$)}。“Query”表示基本检索式；“{ }”中的条件语句对查询内容进行限定；输出语句对输出内容进行限定，并且一个高级检索式中只能有一个输出语句。检索式中被限定的部分需要用“( )”括起来，根据“( )”出现的顺序，可使用“$”符号＋序号取得该

部分内容，进行条件限定或输出限定。第一个“( )”中的成分用“$1”表示，以此类推。如检索式，“(a)(n){len($1)>1；$2=[S_N_名词]；len($2)>1}”表示形容词修饰名词的定中搭配。名词和形容词按“( )”出现的顺序可分别由“$1”“$2”取得。“{ }”中的限定条件表示形容词和名词的长度均大于1，名词限制为“S_N_名词”词表，默认输出“$1”“$2”。通过限制词性、限制长度、限制词表、运用停顿等手段提高抽取结果的准确性。表5是基于BCC的部分抽取结果。

表5 基于大数据的知识抽取平台的部分抽取结果

| 介词结构检索式构成 | 检索结果 | 频次 |
|---|---|---|
| 从(～)上(v){print($1 $2)} | 从根本上解决 | 68 124 |
| | 从表面上看 | 27 763 |
| | 从理论上讲 | 24 708 |
| | 从总体上看 | 18 963 |
| | 从外观上看 | 15 051 |
| | 从理论上说 | 13 003 |
| | 从照片上看 | 12 591 |
| | 从技术上看 | 12 536 |
| | 从本质上说 | 11 943 |
| | 从整体上看 | 10 539 |
| 在(～)上(v){print($1 $2)} | 在电视上看到 | 58 280 |
| | 在网上流传 | 30 695 |
| | 在黑板上写 | 26 443 |
| | 在社会上引起 | 20 501 |
| | 在沙发上睡 | 19 389 |
| 跟(n)(v)W{$2!=[S_形助 V]；print($1 $2)} | 跟党＊走 | 10 249 |
| | 跟妈妈＊说 | 8 570 |
| | 跟人＊说 | 3 558 |
| | 跟老师＊说 | 3 350 |
| | 跟人＊打交道 | 2 935 |
| | 跟朋友＊聊天 | 2 825 |
| | 跟人＊打招呼 | 2 718 |

### 3.2 介词结构搭配检索式的构建

介词结构搭配的检索式主要通过四种手段构建。分别为：(1)通过将介词分类，分别用实例化的介词与动词构成搭配检索式；(2)建立词类搭配表、排他表；(3)运用韵律结构进行长度限制；(4)通过标点W限制等形式手段构建检索式。

以介词短语修饰成分修饰动词的状中搭配检索式为例：首先，将介词进行分类。例如，将介词细分为可以介引地点的介词“P_地点_s_v”此类下，如“朝”“朝着”“顺着”“往”“向”“向着”“沿”“沿着”“由”“在”；然后根据介词分类的属性分别跟名词、代词、方位词、处所词等搭配组成介宾结构；最后跟具体动词分别组成检索式进行知识获取。部分检索式列表如表6所示。

表6 部分检索式列表

| 检索式 |
|---|
| 朝(～)背后(v){mid($1)!=[着 的 在 向 和 著 于 从 往 了 过]；print($1 $2)} |
| 朝(～)边(v){mid($1)!=[着 的 在 向 和 著 于 从 往 了 过]；print($1 $2)} |
| 朝(～)对面(v){mid($1)!=[着 的 在 向 和 著 于 从 往 了 过]；print($1 $2)} |
| 朝(～)里(v){mid($1)!=[着 的 在 向 和 著 于 从 往 了 过]；print($1 $2)} |
| 朝(r)(v)W{$1=[S_R2_指代处所_处]；$2!=[S_形助 V]；print($1 $2)} |
| 朝(r)(v)W{$2!=[S_形助 V]；print($1 $2)} |
| 朝(s)(v)W{$2!=[S_形助 V]；print($1 $2)} |
| 朝着(～)边(v){print($1 $2)} |
| 朝着(～)后面(v){print($1 $2)} |
| 朝着(～)里(v){print($1 $2)} |
| 朝着(～)外(v){print($1 $2)} |
| 朝着(～)外面(v){print($1 $2)} |
| 朝着(f)(v)W{$2!=[S_形助 V]；print($1 $2)} |
| 朝着(r)(v)W{$1=[S_R2_指代处所_处]；$2!=[S_形助 V]；print($1 $2)} |
| 朝着(s)(v)W{$2!=[S_形助 V]；print($1 $2)} |
| 从(f)(v)W{$2!=[S_形助 V]；print($1 $2)} |
| 从(n)(v)W{$2!=[S_形助 V]；print($1 $2)} |
| 从(r)(v)W{$1=[S_R2_指代处所_处]；$2!=[S_形助 V]；print($1 $2)} |
| 从(s)(v)W{$2!=[S_形助 V]；print($1 $2)} |
| 往(_)外(v){print($1 $2)} |
| 往(_)外面(v){print($1 $2)} |
| 往(_)中间(v){print($1 $2)} |
| 往(_)左(v){print($1 $2)} |
| 沿着(_)北(v){print($1 $2)} |

续表

| 检索式 |
|---|
| 沿着(_)两边(v){print($1 $2)} |
| 沿着(_)两旁(v){print($1 $2)} |
| 沿着(_)下(v){print($1 $2)} |
| 在(_)北边(v){print($1 $2)} |
| 在(_)北部(v){print($1 $2)} |
| 在(_)北方(v){print($1 $2)} |

其次，构造词类搭配表或排他表来构造检索式，使检索结果更准确、全面。例如，检索式“从(r)(v)W{$1=[S_R2_指代处所_处];$2!=[S_形助V];print($1 $2)}”中的“S_R2_指代处所_处”根据北大“现代汉语语法信息词典”整理补充而来的，表示可以接处所的指示代词，抽取结果如“从哪儿开始”“从何处着手”“从这里出发”等。

然后，通过韵律结构总结的规律进行长度的限制来构造检索式。如检索式“按照(n)(v)W{$2!=[S_形助V];len($2)=2;print($1 $2)}”动词长度限定为 2 效果较好，抽取结果如“按照规则行事”“按照合同结婚”“按照日子推算”“按照计划执行”等。

最后，通过停顿来构造检索式。例如，检索式“朝着(f)(v)W{$2!=[S_形助V];print($1 $2)}”指该介词结构后紧跟标点的搭配，没有其他的语言句法成分，研究发现使用标点进行限制后的检索结果比不使用标点限制的检索结果准确率要高 10%，这说明在限定语境下搭配效果噪声会相对较小，检索结果会更准确。

表 7 是部分改进后的介词结构做状语修饰动词的状中搭配检索结果。

表 7　部分改进后介词结构做状语修饰动词的状中搭配检索结果

| 检索式 | 释例 | 频次 | 效果评估 |
|---|---|---|---|
| 把(r)(v)W{$2!=[S_形助V];print($1 $2)} | 把车＊开走 | 3 141 | 3 |
| | 把梦＊照亮 | 2 240 | |
| | 把书＊放下 | 2 180 | |
| | 把窗户＊打开 | 1 986 | |
| | 把心胸＊放宽 | 1 959 | |
| | 把火＊扑灭 | 1 889 | |
| | 把时间＊延长 | 1 341 | |

续表

| 检索式 | 释例 | 频次 | 效果评估 |
|---|---|---|---|
| 按(n)(v)W{$2!=[S_形助V];print($1 $2)} | 按常住人口＊计算 | 1 957 | 4 |
| | 按规矩＊办事 | 1 833 | |
| | 按月＊发放 | 1 683 | |
| | 按计划＊推进 | 1 503 | |
| | 按号码查找 | 1 410 | |
| | 按程序办理 | 779 | |
| | 按规定执行 | 664 | |
| | 按原貌重建 | 471 | |
| 归(n)(v)W{$1=[S_N_名词];$2!=[S_形助V];print($1 $2)} | 归父母＊管 | 337 | 3 |
| | 归领导＊使用 | 212 | |
| | 归国家＊承担 | 190 | |
| | 归女方＊抚养 | 123 | |
| | 归学校＊使用 | 97 | |
| 叫(n)(v)W{$1=[S_N_名词];$2!=[S_形助V];print($1 $2)} | 叫我＊起床 | 59 794 | 4 |
| | 叫我＊出去 | 6 275 | |
| | 叫他＊回来 | 5 548 | |
| | 叫我＊吃饭 | 5 451 | |
| 待(r)(v){$1=[S_R2_指代时间_时];$2!=[S_形助V];print($1 $2)} | 待何时＊投稿 | 303 | 4 |
| | 待此＊完成 | 30 | |
| | 待何时＊去 | 16 | |

### 3.3　介词结构搭配的抽取流程

介词结构搭配的抽取流程主要分三步。首先，基于简单介词结构搭配的结构特征构造检索式，将 587 条检索式读入指定文件，以备后期程序调取。其次，利用 BCC 提供的 WebAPI 通过云服务的方式调用已存检索式，抽取搭配知识。最后，将抽取结果进行整合，搭配实例按频次从高往低排序，对于不同的搭配进行不同的阈值限定，最终确定搭配结果，以便后期人工认定。介词结构搭配抽取流程如图 1 所示。

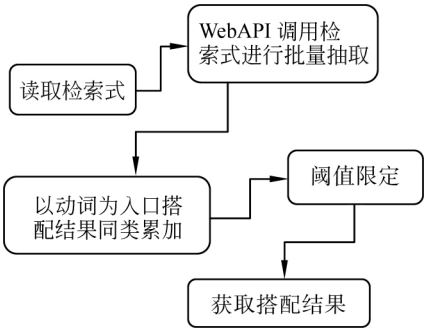


图 1　介词结构搭配抽取流程图

## 4 介词结构检索结果评估及应用

### 4.1 介词结构搭配检索结果评估

本次结果评估是机校和人校合力估值。机器校验主要通过频次及占比排除低频无效的检索式及检索结果;人校主要分两方面:其一是观察检索结果,将效果差的检索式利用形式标记、词长信息或总结规律建立搭配表、排他表等手段改进、删减检索式,其二是人工筛选检索结果。效果评估如表8所示,5指90%以上的正确率、4指70%的正确率、3指50%的正确率、2指30%的正确率、1指10%的正确率。

表8 部分介词结构作状语修饰动词的状中搭配检索式及其评估表

| 修饰成分 | 中心语 | 介词结构检索式 | 效果评估 |
|---|---|---|---|
| 介词结构作状语 | 动词 | 被(n)(v)W{$1=[S_N_名词];$2!=[S_形助 V];print($1 $2)} | 5 |
| | | 在(~)中(v){print($1 $2) | 5 |
| | | 把(n)(v)W{$2!=[S_形助 V];print($1 $2)} | 5 |
| | | 在(r)(v)W{$1=[S_R2_指代处所_处];$2!=[S_形助 V];print($1 $2)} | 4 |
| | | 为(r)(v)W{$2!=[S_形助 V];print($1 $2)} | 4 |
| | | 在(s)(v)W{$2!=[S_形助 V];print($1 $2)} | 4 |
| | | 朝(f)(v)W{$2!=[S_形助 V];print($1 $2)} | 3 |
| | | 照(r)(v)W{$1=[S_R2_指代事物_物];$2!=[S_形助 V];print($1 $2)} | 3 |
| | | 替(n)(v)W{$2!=[S_形助 V];print($1 $2)} | 5 |
| | | 照着(r)(v)W{$1=[S_R2_指代事物_物];$2!=[S_形助 V];print($1 $2)} | 4 |
| | | 向(r)(v)W{$2!=[S_形助 V];print($1 $2)} | 4 |
| | | 直到(f)(v)W{$2!=[S_形助 V];print($1 $2)} | 3 |
| | | 直到(n)(v)W{$1=[S_N_名词];$2!=[S_形助 V];print($1 $2)} | 4 |

通过观察评估结果我们发现,对于有形式化标记和词表限制的检索式,抽取的结果更为准确。常用介词由于使用频率相对较高,其正确率也相应较高。例如,检索式被(n)(v)W{$1=[S_N_名词];$2!=[S_形助 V];print($1 $2)}、在(~)中(v){print($1 $2)、把(n)(v)W{$2!=[S_形助 V];print($1 $2)}的抽取结果比照(r)(v)W{$1=[S_R2_指代事物_物];$2!=[S_形助 V];print($1 $2)}、直到(f)(v)W{$2!=[S_形助 V];print($1 $2)}正确率高20%。跟以往的研究结果相比,本文基于语言学家研究和已有的词典资源,更精准地构造了符合语言学规律的介词结构,主要通过观察检索式结果改进检索式,以迭代检索、机器校验和人工筛选相结合的方法进行知识获取。总共构建了587条检索式,以16 465个核心动词为对象,抽取结构搭配条目1 385 811条。

### 4.2 介词结构搭配结果应用

基于抽取的高质量介宾搭配数据,我们已经对汉语核心动词进行了穷尽性的非核心语义角色认定及相应的知识加工,介词结构的搭配结果为动词语义角色的判定提供了大规模的数据支撑。介宾结构搭配不仅对进一步的搭配知识获取和计算有帮助,也对语义角色标注建模、复述等自然语言处理的句法和语义分析任务有重要的价值,还对语言本体和语言教学研究有重要的意义。

## 5 总结与展望

本文基于大规模语料库构建的介词结构搭配库,主要从介词结构的知识体系建立、抽取及数据评估等角度开展。研究发现,一定的语言规律可以指导知识的获取,然而现代汉语字词句语法、语用歧义无处不在,语言的习用性、动态性和多领域性等特点导致知识获取困难。词性、词长和韵律信息等语言学研究成果可以在很大程度上帮助我们更准确地从语言数据中挖掘出简单的状中搭配,同时基于大规模语料从大数据中获取的搭配资源可以帮助我们找到新的思路,补充或发现新的规律。但是基于大规模语料的状中搭配知识获取也存在缺陷。首先,由于现存的词类划分、词性标注体系不完善,分词、词性标注不准确等问题,检索出的部分语料会有错误,多见于低频部分。其次,现有的语言学研究规律不足以满足自然语言处理的需要,从大数据中进行检索时,抽取的部分搭配噪声较大,不能保证大多数语

料的正确性。最后，由于缺乏上下文，对搭配的认定有局限，不同搭配间会有重叠，不同语体的差异等都会对搭配抽取结果有影响，与现存的研究成果还未进行更精确的对比实验等。

对于存在的问题，未来工作展望如下：首先，针对词性标注体系不完善的部分，我们正在尝试对BCC语料库的动词表进行人工筛选、标注，期望提高后期搭配获取的质量和便于句法语义分析器的部分应用。其次，针对语言学研究规律不彻底的现状，我们可以根据大数据抽取出的结果总结规律，建立搭配表或排他表，再次进行检索，以获得更高质量的数据，同时与现存的研究战果进行对比分析，更准确地评估实验结果的正确率。最后，我们在今后的工作中将改进检索式，用基于分析的方法，构造带有结构属性信息的检索规则进行抽取，使检索语料具有上下文句法结构信息，以期检索结果有质的飞跃。

## 参考文献


[1] 李裕德.现代汉语词语搭配[M].北京：商务印书馆，1998.

[2] 徐敏.由介词·介词结构引起的组合歧义[J].齐齐哈尔大学学报(哲学社会科学版)，1999，04：32-40.

[3] Firth J R. Modes of meaning in papers in linguistics [M]. Oxford：Oxford University Press，1957.

[4] Sinclair J. Corpus concordance collocation[M]. Oxford：Oxford University Press，1991.

[5] Singleton D. Language and the lexicon：An introduction[M]. London：Arnold，2000.

[6] Kjellmer G. Some thoughts on collocational distinctiveness in recent developments in the use of computer corpora in English language research[J]. Costerus，1984，45：163-171.

[7] Iison M，Benson E，Benson R. The BBI combinatory dictionary of English：A guide to word combinations [M]. Amsterdam，John Benjamins Publishing，1986.

[8] Xu R，Lu Q，Wong K F，et al. Building a Chinese collocation bank[J]. International Journal of Computer Processing of Languages，2009，22 (01)：21-47.

[9] 邢公碗.语词搭配问题是不是语法问题？[J].安徽师范大学学报(人文社会科学版)，1978，(4)：77-84.

[10] 林杏光.张寿康先生与词语搭配研究[J].首都师范大学学报(社会科学版)，1995，(1)：59-63.

[11] 林杏光.词语搭配的性质与研究[J].汉语学习，1990，(1)：7-13.

[12] 林杏光.论词义分类和词语搭配[J].中国人民大学学报，1991，(5)：77-82.

[13] 胡清国，高倩艺.词语搭配与对外汉语教学[J].语言与翻译，2017，(04)：58-61.

[14] 张寿康，林杏光.学生常用词语搭配词典[M].石家庄：河北少年儿童出版社，1989.

[15] 张寿康，林杏光.简明汉语搭配词典[M].福州：福建人民出版社，1990.

[16] 张寿康，现代汉语实词搭配词典[M].北京：商务印书馆，1999.

[17] Choueka Y，Klein S T，Neuwitz E M，et al. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus[J]. Journal for Literary and Linguistic Computing，1983，4 (1)：34-38.

[18] Church K W，Hanks P. Word association norms，mutual information and lexicography[J]. Computational Linguistics，1990，16 (1)：22-29.

[19] Kilgarriff A，Rychly P，Tugwell D，et al. The sketch engine[J]. Information Technology，2004：105-116.

[20] Huang Ch R，Kilgarriff A，Wu Y Ch，et al. Chinese Sketch Engine and the extraction of grammatical collocations[C]//Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing，2005：48-55.

[21] Hu R F，Chen J，Chen K. The construction of a Chinese collocational knowledge resource and its application for second language acquisition[C]//Proceedings of COLING 2016，the 26th International Conference on Computational Linguistics：Technical Papers，2016：3254-3263.

[22] Smadja F. Retrieving collocation from text：Xtract [J]. Computational Linguistics，1993 (01)：143-177.

[23] 李洪政，晋耀红.汉语介词短语自动识别研究综述[J].中文信息学报，2017 (02)：6-15.

[24] 温苗苗，吴云芳.基于SVM融合多特征的介词结构自动识别[J].中文信息学报，2009，23(5)：19-24.

[25] 鉴萍，宗成庆.基于双向标注融合的汉语最长短语识别方法[J].智能系统学报，2009，4(05)：406-413.

[26] 卢朝华，黄广君，郭志兵.基于最大熵的汉语介词短语识别研究[J].通信技术，2010，5：181-183，186.

[27] 干俊伟，黄德根.汉语介词短语的自动识别[J].中文信息学报，2005，19(4)：17-23.

[28] 奚建清，罗强.基于HMM的汉语介词短语自动识别研究[J].计算机工程，2007，3：172-173，182.

[29] Li H Q，Huang Ch N，Gao J F，et al. Chinese chunking with another type of spec[C]//Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing，2004：24-26.

[30] 霍亚格，黄光君.基于最大熵的汉语短语结构识别方法[J].计算机工程，2011，37(16)：206-208，211.

[31] 胡思磊.基于CRF模型的汉语介词短语识别[D].大连：大连理工大学硕士学位论文，2008.

[32] 朱丹浩，王东波，谢靖.基于条件随机场的介宾结构自动识别[J].现代图书情报技术，2010，(Z1)：79-83.

[33] 卢朝华，徐好芹，王玉芬.基于语义分析的汉语介词短语识别方法研究[J].电脑与电信，2012，3：46-48.

[34] 张杰.基于多层CRFs的汉语介词短语识别研究[D].大连：大连理工大学硕士学位论文，2013.

[35] 张灵.基于层叠条件随机场的汉语介词短语识别研究[D].沈阳：沈阳航空航天大学硕士学位论文，2013.
[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，2016，3(1)：93-109，118.
[37] 李晓琪.现代汉语虚词讲义[M].北京：北京大学出版社，2005.
[38] 马杜娟，郑通涛.现代汉语常用介词语块研究[M].广州：世界图书出版广东有限公司，2016.
[39] 傅雨贤.现代汉语介词研究[M].广州：中山大学出版社，1997.
[40] 陈昌来.汉语“介词框架”研究[M].北京：商务印书馆，2014.
[41] 俞士汶.现代汉语语法信息词典详解[M].北京：清华大学出版社，1998.

邢丹(1992—)，硕士研究生，主要研究领域为句法语义分析、语言资源建设、搭配库构建。
E-mail：xingdan1@126.com

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

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