文章编号：1003-0077(2020)09-0019-09

三元搭配视角下的汉语动词语义角色知识库构建

王诚文[1]，钱青青[1]，荀恩东[1]，邢 丹[1]，李 梦[1]，饶高琦[1,2]

(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：动词语义角色一直是国内外语言学界研究的重点和难点。在自然语言处理领域，相关的语言资源也在逐步构建。对于汉语而言，国内大部分工作集中在语义角色标注上。该文创造性地提出了一种三元搭配的动词语义角色知识表征形式，并在前人研究的基础上，提出了一套语义角色分类体系。在该体系指导下，对汉语动词进行了穷尽式的语义角色认定及相关知识加工，以构建汉语动词语义角色知识库。截至目前，该工程考察了 5 260 个动词，加工了语义角色及引导词的动词数量为 2 685 个，加工认定语义角色 4 307 个。

关键词：三元搭配；语义角色；实例化；引导词

中图分类号：TP391 **文献标识码**：A

Construction of Semantic Role Bank for Chinese Verbs from the Perspective of Ternary Collocation

WANG Chengwen[1], QIAN Qingqing[1], XUN Endong[1], XING Dan[1], LI Meng[1], RAO Gaoqi[1,2]

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

Abstract: The research on semantic roles has always been a significant challenge in the field of linguistics. Some resources on the semantic relations have been constructed; however, most of the domestic researches on Chinese word semantic relations focuses on the labeling. This paper proposes a novel structure, the ternary collocation, to describe the semantic relations with verbs at the core. The paper also puts forward a semantic role classification scheme, under which a semantic role bank for Chinese verbs is constructed. All the verbs involved are exhaustively identified for the possible semantic roles and other related knowledge annotation. Altogether 5,260 verbs are collected, among which 2,685 verbs are assigned with 4,307 semantic roles as well as the guiding word.

Keywords: ternary collocation; semantic role; instantiation; guiding word

0 引言

动词一直以来是语法研究的中心及重心。句子的理解也主要围绕动词性成分与其周围所支配从属成分的句法语义关系展开。动词语义角色一直以来都是中外语言学领域着重关注的研究对象。与此同时，在自然语言处理领域，动词语义角色划分体系及相关知识库的构建工作，也是许多学者持续关注的研究重点。

构建以动词为中心的语义角色知识资源对于语言本体、教学及自然语言处理研究有着重要意义。一方面，该资源能够为语言本体的动词句法语义研究提供数据支撑。另一方面，于自然语言处理而言，其对语义角色标注模型构建、复述技术及深层句法语义分析任务等都有着重要的价值。结合具体应用任务，该资源能够为信息抽取任务提供引导性知识，解决歧义问题。在对话及聊天系统任务中，核心问

收稿日期：2019-09-09 定稿日期：2019-10-19

基金项目：国家社会科学基金(16AYY007)；北京市语言资源高精尖创新中心项目(TYR17001)；北京语言大学校级项目(教育部高校科研基本业务费 18WT03)；北京语言大学研究生创新基金(20YCX144)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

题是能够对于用户问句进行理解，并在知识库中进行答案搜索匹配，最后生成合适的句子予以反馈。以动词为核心的框架性语义角色知识能够有效呈现句子的核心框架，有效赋能句子的理解和生成。

本文创造性地提出一种三元搭配的动词语义角色知识表征形式，并在前人研究的相关成果上，提出了一套语义角色分类体系。在该体系指导下，以大规模语料为支撑，对汉语动词进行穷尽性的语义角色认定、语义角色实例提取及相关知识加工，以构建汉语动词语义角色知识库。

本文组织结构如下：第1节对动词语义角色体系及知识加工相关研究现状进行综述；第2节介绍了本研究确立的动词语义角色知识表征形式及语义角色体系；第3节是对目前知识库构建进展情况的一个总结；第4节对知识库的呈现形式及其应用进行了介绍；最后对全文工作进行了总结。

1 相关研究

1.1 动词语义角色体系相关研究

国外，动词语义角色分类体系的相关研究按照其术语体系的不同可以分为三种主要类型。一种是特思尼耶尔在《句法结构基础》中提出的动词配价研究，根据动词所支配名词成分，将动词分为一价动词、二价动词和三价动词。第二种主要是以菲尔莫为代表的“格”语法的研究。菲尔莫在其著作《“格”辨》(The case for case)中提出了格语法，主要包括6种格：施事格、工具格、客体格、处所格、承受格和使成格。有学者[1]在对菲尔莫研究工作进行总结基础上提出，菲尔莫后期的格数量扩展到13个，主要为施事格、工具格、客体格、处所格、承受格、感受格、源点格、终点格、时间格、行径格、受益格、伴随格和永存格。还有一种工作聚焦在语义角色的分类研究上，宾西法尼亚大学构建PropBank时，定义了20多个语义角色，其中核心的语义角色为Arg0～Arg5六种。

国内动词语义角色的研究工作主要聚焦在语义角色认定及分类上。袁毓林[2]界定了汉语中常见的17种论元角色，并给出了其宽泛的语义定义，主要包括：施事、感事、致事、主事、受事、与事、结果、对象、系事、工具、材料、方式、场所、源点、终点、范围和命题。鲁川等[3-4]则将格关系分为7大类，22小类。朱晓亚[5]界定了14种语义角色关系。孟琮[6]在《动词用法词典》中从动宾结构中名词宾语类别出发，划分出了14种类别，其中提出了一种“杂类”的语义分类。后续的学者[7-9]也在语义角色数量的认定上有着不同的认识，在格的性质认定及分类上做了一系列研究工作，总体上呈现出一种“大框架稳定，小细节灵活”的面貌。然而，对于语义角色的分类研究，鲜有结合具体应用场景的落地化考虑。服务于知识库构建，国内相关研究机构也定义了自己的语义角色标注体系。山西大学建设的汉语框架语义知识库(CFN)[10]定义了31种常用的动词周边语义角色。董振东先生[11]构建的知网则提出了事件内部语义关系总计为83类。

从调研情况看来，国内外学者都围绕动词构建了相应的语义角色知识体系。各家的语义角色分类数量有多有少，在具体语义角色术语的使用上有交叉也有差别。不同体系之间术语和内涵的对应呈现出“名同实异”及“名异实同”的特点。然而，过于细粒度的语义角色划分体系中部分语义角色之间的区分性不是特别显著，给知识库构建者增加了区分负担，不利于大规模语言资源的快速构建。同时，语义角色体系的构建需要面向具体的应用场景，而非仅追求学理逻辑上的严密性，如此才能够做到与领域问题的有效嫁接。

1.2 动词语义角色相关知识库构建现状

动词语义角色知识库构建的相关研究可以根据其知识的形式分为两种主要类型。一种是标注资源。国外较早进行语义角色知识库构建的工作应属U.C. Berkeley[12]基于框架语义学理论开发的FrameNet。该库选取英语国家语料库进行标注，重点在于刻画每个谓词的语义框架。然而，该工作定位为一种标注工作，采用数据标注的方式来承载框架语义知识，囿于标注语料规模及语体限制，知识库的全面性及应用上的鲁棒性可进一步提升。国内的语义角色标注树库主要有北大依存树库[13]、哈工大依存树库[14]、CPB(Chinese Proposition Bank)[15]、北大中文网库等。另一种是词典资源，围绕具体动词进行语义角色刻画。胡佛汉顿大学研究组[16]基于语料库模式分析技术，建立了PDEV(Pattern Dictionary of English Verbs)，对英语中动词出现的常用语义模式进行了总结归纳。该工作计划建立5 369个英文动词模式数据库，目前已经完成1 364个动词模式和例句库的构建工作。北京大学中文系袁毓林主持建设的《北京大学现代汉语实词句法语

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

义功能信息词典》[17]是一个电子化的语言知识资源，知识内容主要是现代汉语常用形容词、动词和名词的句法功能、语义角色及其组配方式、主要句型及其典型例句。常用形容词 3 000 个，4 000 个义项条目；常用动词 6 000 多个，8 000 个义项条目；常用名词 1 万多个，1.2 万个义项条目，该资源目前还没有公开开放。鲁川和林杏光先生编撰了《动词大词典》[4]，对常用动词的 3 000 多义条能够支配的语义角色进行了刻画。该词典目前仍旧是一种辞书形式，没有转换成系统的可服务于自然语言处理需求的形式化语言资源。

目前国内外动词语义角色相关知识库构建工作已经取得很大进展，但在语义角色知识的知识规模和表征形式上都有待进一步提升。以标注数据来承载语义角色知识的形式，由于受到语料标注规模及领域适用性的限制，会产生重复标注高频动词而极少刻画低频动词语义角色的问题，且对人力成本的高消耗导致了语义角色标注树库规模较小的现状。在词典资源构建方面，大部分以语义角色或对应语义类与动词的组合配位方式来表征具体知识，例如，施事＋把＋受事＋v，human＋吃＋food。对语义角色对应具体实例化知识缺乏有规模的获取。同时，从资源构建规模来说，语义角色标注树库规模通常维持在几万句左右，对动词及其语义角色的覆盖面较小。本研究初步以《现代汉语词典》(第 5 版)的 15 891 个双音节动词为对象，从大数据角度出发，进行语义角色认定及语义角色对应实例的加工工作。

2　动词语义角色知识表征形式及分类体系构建

2.1　动词语义角色知识表征形式

当前面向自然语言处理的动词语义角色资源构建工作，大部分以数据标注形式来呈现具体知识，部分学者则以语义角色或对应语义类与相应动词的组合配位方式来对具体知识进行表征。

本研究在大数据背景下，注重对语义角色实例化知识的刻画，同时强调语义角色形式化标记的语义凸显作用。大部分语义角色能够通过“格标记”或与动词的相对位置来显化，“格标记”“语义角色”“动词”的三元搭配能够呈现出句子基本语义结构模式；在实际语言中，句子语义模式更为多变，但大多数仍可以通过基本模式的变形得到；且由“动词”“格标记”“语义角色”三元组合构成的知识表征形式形式化特征强，易提取。鉴于此，本文提出了一种三元搭配的动词语义角色知识表征形式，以期能够涵盖实际语言中的语义角色。

三元搭配的语义角色知识表征形式：由典型引导词①、核心谓词及谓词相关语义角色构成的跟动词密切相关且高频的具备强搭配性的形式、意义自足的语义表征组合。三元搭配有“引导词＋语义角色＋动词”及“动词＋引导词＋语义角色”两种形式。

通常情况下，引导词由介词充当，也可由动词充当，谓词相关语义角色一般为名词性实体概念。同时，考虑到语义角色的通用性及其结构特征，本研究仅提取其中心语部分。因此三元搭配的主要表征形式如表 1 所示。

表 1　三元搭配动词语义角色知识示例

引导词＋语义角色＋动词	动词＋引导词＋语义角色
向_记者_透露	上涨_至_80%
用_混凝土_建造	移交_给_下属
以_肩膀_支撑	走路_到_学校
往_南方_奔跑	相比_于_电影

按照引导词与动词的结合程度可将“引导词”“语义角色”“动词”的三元组合分为几种类型：

(1) 引导词和动词相关度较高，但引导的语义角色不固定；

(2) 引导词和动词低相关，在实际语料中可与较多类别动词搭配；

(3) 引导词和动词结合不紧密，多数通过变形放在句首，呈现为“句饰语”的形式；

(4) 由单个动词所组成的三元搭配语义上不自足，形式上不完整。

因此在概念界定和后续工作中，本研究着重突出三元搭配的两个特征，一个是“动词密切相关性”，另一个是“强搭配性”。

“动词密切相关性”在这里主要指三元搭配的语义组合与动词相关。引导词的确定是跟具体动词有关的，该引导词一般只能和该动词或该动词的同类动词相搭配。

通过图 1 和表 2，以引导词“跟”“把”“被”“在”

① 引导词，对应到格语法中，主要是指格标记。在汉语中，动词语义角色主要由介词牵引出来，这里也包括部分动词。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

的对比，可以发现不同三元组内部的搭配能力有高低，在朱晓亚提出的42个动词小类中，“把”“被”“在”相比“跟”能够引导更多的动词类别，在与具体动词的搭配上较“跟”而言更为泛化，跟动词相关度相对较差。A组的例子中，通过引导词的介引作用，牵引出来了后边动词所表示动作行为涉及到的对象成分。具体引导词的选择是由动词所表示的含义限制决定的。本文围绕具体动词进行语义知识构建时，主要考虑的是A组那样与动词密切相关的引导词及语义角色实例知识的构建。

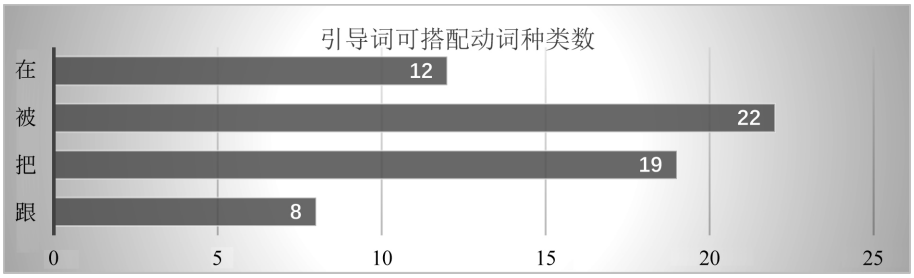

图1 引导词与不同类别动词搭配性分布

表2 引导词与不同类动词搭配示例

A跟	B把	C被	D在
—	把_电视_关掉	被_老师_关掉	在_上午_关掉
—	把_学生_责骂	被_主人_责骂	在_早上_责骂
—	把_头发_吹乱	被_大风_吹乱	在_刚才_吹乱
跟_别人_诉苦	—	—	在_时机_诉苦
跟_同学_转告	把_某事_转告	被_朋友_转告	在_当天_转告

“强搭配性”主要是指“引导词”“语义角色”及“动词”能够组成形式上完整、意义上自足且与动词连接紧密的三元搭配。例如，“跟_自己人_协商”“送_给_自己人”。

表3给出了搭配示例。对比A、C两组例子，可以发现，A组的三元搭配组合更为松散，体现为大多能将“引导词＋语义角色”通过变形提到句首，成为“句饰语”，整体上修饰整个句子。而C组中三元搭配较为紧密，一般不会将语义角色提前。

表3 搭配强弱对比示例

A	B	C
对于这件事情，老师发表了看法	为×打	为×打水
就本问题，全组展开了探讨	与×发生	与×发生冲突
关于这件事，我没什么看法	跟×过	跟×过不去

对比B、C两组例子，可以发现B组的例子在语义理解上不完整，而C组的例子，动宾结构整体与引导词形成搭配对时形式和语义上都是自足的。本文在动词知识库构建中，注重三元搭配对的形义完整性。

除此之外，语义角色的认定中，还坚持“高频”准则。“高频”主要强调“引导词”“语义角色实例”及“动词”构成的三元搭配在大规模语料中的高频稳态出现特征。动词支配语义角色的确定问题符合齐普夫定律(Zipf’s Law)。在词汇层面，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。类比而言，动词的极少数高频语义角色数(type)已经占了该动词语义角色出现次数(token)的绝大部分。鉴于此，本文在进行动词语义角色刻画及引导词确立时，参考其在大规模语料中的出现频次。

2.2 构建动词语义角色分类体系

无论是汉语本体研究还是面向自然语言处理的研究，在动词语义角色的分类数量及术语内涵界定上，都自成一家之言。过于细粒度的语义角色区分，一方面对基于该体系的知识库构建工作带来较大负担；另一方面，在跟具体应用任务结合的时候，细粒度语义角色在与实际“槽”的对应上难度较大。

本文语义角色分类体系的构建，是在结合汉语语义角色分类的代表性工作——袁毓林[2]、鲁川[3]和朱晓亚[5]的基础上整合而来的。具体的体系构建过程及最终体系如下：

首先，对三位学者构建体系的术语名称与具体内涵做了比较研究。具体对照如表4所示。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 4　语义角色对照表

作者及分类数量		鲁川、林杏光(22)		袁毓林(17)	朱晓亚(14)
语义角色	核心语义角色	主体	施事	施事	施事
			当事	感事	经事
				主事	系事
			领事		起事
		客体	受事	受事	受事
			客事	对象	涉事
			结果	结果	成事
		邻体	与事	与事	当事
			同事		共事
			基准		
		系体	系事	系事	止事
			分事		
			数量		
	外围语义角色	工具		工具	
		材料		材料	
		方式		方式	
		范围		范围	
		时间			
		处所		场所	
		方向		源点	
				终点	
		依据			
		原因			
		目的			
	独立角色			命题	补事
				致事	准客事
					使事
					位事

如表 4 所示，对于三位学者的语义角色体系，根据术语对应内容相同或相近的原则，做了一致对应。从表中发现以下几点情况。首先，各家体系在术语使用上呈现出“名同实异”及“名异实同”或“名异实近”的特点，且以后者为主。朱晓亚体系中的“系事①”和鲁川体系的“当事”及袁毓林的“主事”所指内涵一致，但术语名称却相差较大，与袁毓林和鲁川体系中的“系事②”所指形成巨大的反差。其次，袁毓林和鲁川在术语使用上较为一致，相同或相近术语对应内涵基本一致。两者共同使用的术语数量为 9 个。内涵一致的术语对应对为 15 个。可见两者的语义角色划分体系一致性较高，利于两者体系的融合。最后，朱晓亚体系的术语相对更加抽象，前两位的术语可接受度及可理解度相对较高。

基于上述分析，遵循“对立互补原则”，以鲁川的术语体系基础，对鲁川和袁毓林体系进行了有效的融合，形成新融合知识体系。该体系一共包括 15 个语义角色，为适应具体应用场景对于外围语义角色精细度要求高的需求，在核心论元角色分类上从宽，在外围语义角色上从严。具体如表 5 所示。

表 5　融合后语义角色分类体系

融合语义角色体系		
核心语义角色	外围语义角色	
主体	工具	范围
客体	材料	时间
邻体	方式	处所
	依据	原因
	源点	目的
	终点	数量

该融合体系在继承中发展。一方面，对各家体系中本身区分度较差的语义角色做进一步提升，以减轻语言工程开展中的区分负担。另一方面，考虑到具体应用场景中对外围语义角色需求精细的特点，在外围语义角色上尽量做到详细、覆盖面广。在语义角色数量上，适中的语义角色数量能够便于大规模语义角色标注或知识构建工作的开展，同时也便于机器学习模型的构建。

3　知识库构建

3.1　加工对象

本研究加工动词主要选取《现代汉语词典》第 5 版(以下简称现汉)中的动词为穷尽式考察对象。现汉中双音节动词 15 891 个，单音节动词 1 564 个。

① 朱晓亚将“系事”界定为和性状动词所联系着的主体动元，是性状的系属者，如“他很聪明”和“房子倒塌了”中的“她”和“孩子”便是“系事”。

② 袁毓林和鲁川中的系事，指事件类别的主体、身份或角色，如“我有一本书”和“我是中国人”中的“我”便是。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

单音节动词的多义问题较为突出，而可以通过引导词、语义角色实例及动词的三元搭配对双音节动词中较少的多义现象进行有效的消歧，因此目前考察对象主要集中于双音节动词，以双音节动词的词条为加工对象。本研究基于BCC(北京语言大学现代汉语语料库)[18]，得到了现代汉语词典中双音节动词的词频表，由高到低对动词进行过滤考察。

在具体动词的语义角色确定上，考虑到同时对每个动词进行15个语义角色的考察工作的负担，该工程分两步走。首先加工与动词密切相关的7个语义角色，即邻体、工具、材料、方式、依据、源点和终点。后期，对其他语义角色进行规则化处理。

3.2 加工流程

在动词支配语义角色及引导词的确定上，本部分研究工作充分利用了BCC的大数据检索处理能力。BCC总规模达数百亿字，能够有效地为语义角色及引导词的认定提供定量数据的支持。

具体的加工流程归纳为：基于词典释义的语义角色提升、基于BCC的量化三元搭配知识抽取、具体动词语义角色及引导词确定、数据抽取和校对入库。详细的加工步骤阐释如下。

步骤1 基于词典释义的语义角色提升

在该步骤，主要依据现汉的释义进行实例到语义角色类型的提升。例如，动词【装载】在词典中的释义为“用运输工具载(人或物)”。将该解释中的“运输工具”实例提升为动词“装载”的“工具”语义角色，并确立其典型引导词为“用”。词典对动词语义的刻画是比较稳固和全面的，对于能够从释义中提升出来的语义角色，首先做了认定。

步骤2 基于BCC的量化三元搭配知识抽取

依托BCC强大检索能力，构建语义角色引导词表①，制定动词三元搭配知识抽取规则，为具体动词的语义角色和语义角色相应引导词的认定提供量化参考数据。如检索式：(p)＊n 看齐{$1=[P_邻体_b]}，能够有效地从大规模语料中抽取出来跟动词“看齐”搭配的邻体引导词及邻体语义角色对应实例。在BCC中的一个子语料上进行检索得到结果如表6所示。

本工作结合每个动词都进行了7种动词语义角色三元搭配实例的抽取，当某类语义角色的抽取数据条数为零时，则倾向于认定该动词不具备该语义角色。否则，对于抽取实例，对其高频数据进行观察，结合主观判断，进行语义角色及相应引导词的认定。

表6 检索式示例

向＊标准看齐 50	向＊水平看齐 45
向＊人看齐 44	向＊大学看齐 33
向＊理念看齐 25	向＊学院看齐 20
向＊企业看齐 11	向＊高校看齐 5

步骤3 具体动词语义角色及引导词确定

参照语义角色体系，围绕具体某一个动词，采用内省与数据分析相结合方式，对于具体的某一个动词能够经常支配哪些语义角色及经常搭配引导词给出明确认定。在该部分结合的分析数据主要是指在步骤2中获取到的带有频次信息的数据。表7给出了动词语义角色及引导词认定示例。

表7 动词语义角色及引导词认定示例

动词	邻体	方式	源点	终点
道歉	向、给、对			
出发		以	从、自	向、往
来往	与、跟、和			
比赛	与、跟、同			
诉苦	向、跟、给			

步骤4 数据抽取

在上一步工作的基础上，基于自研发的语言大数据处理平台，构造更为严密的相应知识抽取检索式，从大规模语料中抽取实例化知识，并根据概率分布信息进行排序。例如，动词“诉苦”具备语义角色“邻体”，其相应的引导词为“向、跟、给”。可以分别构建“向＊n 诉苦”“跟＊n 诉苦”“给＊n 诉苦”三个检索式，对于其抽取出来核心名词n进行求交集操作。

步骤5 校对入库

对于抽取出来的高频数据进行进一步校对，并存入数据库中。

3.3 知识库建设进展

工程自开展以来，穷尽式地过滤动词数目为5 260个，刻画了语义角色及引导词的动词数量为

① 语义角色引导词表，主要是参考介词语义类与语义角色对应关系的基础上构建的以帮助三元搭配抽取的词表。如“邻体类”引导词包括“和、跟、与、同、连、对、为、给、替、向、管、问、给、拿、朝”。具体词表见附录1。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 685 个，共计 4 307 个语义角色。截至目前，工程主要是进行步骤 1～3 的工作。在本小节对知识库加工进度、数据加工一致率以及基于加工数据库的动词及语义角色分布情况进行详细说明。

该知识库加工工作以周（期数）为单位，目前进行了 14 周（期）的加工工作。其中 1/2/3/6 期为加工培训阶段，其具体加工安排情况在表 8 中予以介绍。具体每期加工情况如表 8 所示。

表 8　动词加工量进展

期数	组数	天数	平均量/天	总量
L4	2	5	50	500
L5	4	5	30	600
L7	4	5	30	600
L8	4	4	20	320
L9	4	3	30	360
L10	4	4	30	480
L11	4	5	30	600
L12	4	5	30	600
L13	4	5	30	600
L14	4	5	30	600

表 8 中，每组为两两加工，不一致的地方由第三方进行审核认定，以尽可能保证动词语义角色加工的质量。“平均量/天”代表进行穷尽性过滤考察的动词数量。截至目前已经穷尽性过滤动词5 260 个。

该知识库的加工一直采用“两两加工、第三方审核”的形式进行。具体的加工一致率统计如图 2 所示。

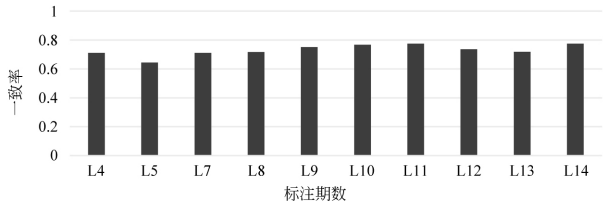

图 2　标注一致率情况

这里的加工一致率主要是指动词语义角色认定上的一致率，即对于某动词的某个语义角色，加工者 A 和加工者 B 同时给予认定的一致率。从图 2 中可以发现，动词语义角色加工认定的一致率普遍维持在 70% 以上。10 期的加工平均一致率为 73.02%。

在穷尽性加工 5 260 个动词后，加工了语义角色的动词数量为 2 685 个。在表 9 和表 10 中，对语义角色与动词的分布情况做了统计。

表 9　语义角色认定数与动词数对照表

语义角色认定数	动词数量
6	3
5	17
4	75
3	282
2	750
1	1588

如 3.1 加工对象一节明确介绍所示，该知识库主要围绕与动词密切相关的 7 种语义角色进行加工认定。在已经给予语义角色认定了的 2 685 个动词中，有三个动词认定了 6 个语义角色，为“开车”“按摩”“发送”。近 60% 的动词只认定了一个语义角色，这一数据符合幂律分布的特点。在进行数据分析后，发现语义角色认定数量大于或等于 3 的动词大部分为具有实在动作行为义的动词，或者说倾向于具备实在含义。

表 10　语义角色与认定动词数对照表

语义角色	认定该语义角色动词数
邻体	1351
方式	1179
依据	711
工具	480
终点	262
源点	260
材料	75

2 568 个动词中，“邻体”“方式”“依据”是认定动词数量排前三位的语义角色。认定了“材料”的动词数量出现最少，主要是因为“材料”语义角色需要其支配动词语义含义较为实在，如“装修”“建造”等。

4　知识库呈现形式及应用

4.1　知识库呈现形式

知识库最终将呈现为“动词—语义角色—引导词—语义角色实例”形式，针对每一个动词将给出其

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

所有的语义角色，针对该动词的每一个语义角色将会给出引导词及其在大规模语料中所对应的具体高频实例，现以动词“抗争”为例简要说明。

如表11所示，对于动词“抗争”，从大规模语料出发认定其有“邻体”和“方式”两类语义角色，“邻体”中对应有“与”“跟”“同”介引的各类实例，方式中有“用”“以”所介引的实例。

表11 动词“抗争”动词语义角色知识示例

动词	语义角色	引导词	语义角色实例
抗争	邻体	与	命运、病魔、癌症、赖床、死神、疾病、苦难、抑郁症……
		跟	命运、病魔、人、妈、NR、蚊子、现实、台风、病毒、东西……
		同	命运、病魔、风暴、疾病、贩子、命运、癌症、死神、势力……
	方式	用	方式、命、生命、生命、暴力……
		以	命、方式、生命、武力、命运、理、理性……

相较于以往用标注数据或字典资源承载语义知识的形式，本知识库能够更大面积地覆盖现代汉语中所有动词的语义角色，这是其他语义资源所不具备的。本知识库也更强调在实际语言应用中动词所支配语义角色的呈现，如上文中动词“抗争”，就其“邻体”语义角色而言，理论上和“与”类引导词同类的介词“和”在实际语料中非常低频出现，因此在知识库中不予呈现，以减少人的学习成本、机器的搜索空间和决策空间。本知识库能够针对动词的语义角色给出具体的、可扩充的实例。如“抗争”，在知识库中共有287个邻体的实例，在具体语境中能够以中心语的形式涵盖绝大部分具体例子。这样可以看出本知识库能够给出更大范围的标注语料，更好地服务于自然语言处理的各个任务。

4.2 知识库应用

4.2.1 面向自动句法语义分析的汉语动词语义角色知识库

自然语言理解的词语在很大程度上是对句子的论元结构进行分析呈现。可以将该动词语义角色知识，以三元搭配的形式注入到句法语义分析器中，在“大词库，小规则”的指导下，通过符号计算对句子的论元知识进行分析。

4.2.2 面向本体研究的汉语动词语义角色知识库

从业已构建的知识库来看，认定了“邻体”语义角色的动词数量最多。统计观察后发现，具备“邻体”类语义角色的动词可以分为以下几类，如表12所示。

表12 “邻体”语义角色认定动词的种类

动词类别	动词	作谓语常用格式
框式动词①	着想	为＋n＋着想
	结婚	跟＋n＋结婚
	作对	跟＋n＋作对
三价动词	请教	向＋某人＋请教
	授予	向＋某人＋授予
	叮嘱	向＋某人＋叮嘱
动宾结构动词（框式动词除外）	对话	跟＋某人＋对话
	赋能	为＋某事＋赋能
	助力	为＋某事＋助力

通过对比可以发现，表12中的动宾结构动词的句法语义特点具备很大独特性。在实际语料，尤其是标题句中能够看到该类动词后边直接加宾语的例子，如：“赋能语言信息处理产业”“助力乡村经济建设”。在本知识库的支撑下，可以对知识库中刻画了邻体语义角色的动词做进一步考察，确定动宾结构动词的具体数量，能够利用该部分数据进行句法语义特征的研究分析。

5 结语

动词语义角色知识体系及相关知识库构建工作一直是国内外学者关注的研究方向。然而能够以大数据知识为支撑，穷尽式对汉语动词进行语义角色加工及实例化语义角色知识获取的工作却鲜有。本文的尝试性工作为该路线做了有益的探索。

于本体语言研究而言，该部分动词语义角色知识库能够为本体研究提供大量语言现象，三元搭配的数据能够促进本体的搭配研究工作。

从自然语言处理领域来看，汉语具备强意合型特征，给自然语言处理的句法和语义分析带来很大困难。以动词语义角色引导词为形式化标记，以海量规模语料为数据支撑，获取大规模实例化动词语义角色三元搭配知识（引导词、语义角色实例和动

① 框式动词是沈萍[20]率先提出的概念，主要指在句法上表现为一种框式结构，作谓语时必须与介词搭配才能使用，在语义上一般呈现出“指向性”、“单相向”、“主动性”的语义特征。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

词），能够为基于符号运算的语义分析提供知识引导。同时，以该部分知识为数据，做向量化嵌入表征，能够基于此数据做参数计算，以构建语义角色自动识别模型或为述语复述技术提供数据支撑。后续本团队也将寻求以一定方式与学术界共享资源，促进自然语言理解工程应用的发展。

参考文献

[1] 冯志伟. 从格语法到框架网络[J]. 解放军外国语学院学报，2006(03)：1-9.

[2] 袁毓林. 论元角色的层级关系和语义特征[J]. 世界汉语教学，2002(03)：10-22，2.

[3] 鲁川，林杏光. 现代汉语语法的格关系[J]. 汉语学习，1989(05)：11-15.

[4] 鲁川，林杏光. 动词大词典[M]. 北京：中国物资出版社，1994.

[5] 朱晓亚. 现代汉语句模研究[M]. 北京：北京大学出版社，2001.

[6] 孟琮. 汉语动词用法词典[M]. 北京：商务印书馆，2012.

[7] 李临定. 现代汉语句型[M]. 北京：商务印书馆，1986.

[8] 傅雨贤，刘街生. 现代汉语语法学[M]. 广州：中山大学出版社，2002.

[9] 沈阳，郑定欧. 现代汉语配价语法研究[M]. 北京：北京大学出版社，1995.

[10] 郝晓燕，刘伟，李茹，刘开瑛. 汉语框架语义知识库及软件描述体系[J]. 中文信息学报，2007，21(05)：96-100，138.

[11] 董振东，董强，郝长伶. 知网的理论发现[J]. 中文信息学报，2007，21(04)：3-9.

[12] Collin F Baker, Charles J Fillmore, John B Lowe. The Berkeley FrameNet Project [C]//Proceedings of the COLING2ACL. Montreal，Canada, 1998, 1-5.

[13] 邱立坤，金澎，王厚峰. 基于依存语法构建多视图汉语树库[J]. 中文信息学报，2015，29(03)：9-15.

[14] 陈鑫. 基于主动学习的汉语依存树库构建[D]. 哈尔滨：哈尔滨工业大学硕士学位论文，2011.

[15] Nianwen Xue. A Chinese lexicon of roles and senses [J]. Journal of Language Resources, 2006, 40(3-4): 395-403.

[16] Hanks P. Web page (pilot study for a much larger project): Pattern Dictionary of English Verbs (PDEV)[DB/OL]. 2010-01-01. https://uwe-repository.worktribe.com/output/985376.

[17] 袁毓林，卢达威. 怎样利用语言知识资源进行语义理解和常识推理[J]. 中文信息学报，2018，32(12)：11-23.

[18] 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(01)：93-109，118.

[19] 中国社会科学院语言研究所词典编辑室. 现代汉语词典(第 5 版)[M]. 北京：商务印书馆，2005.

[20] 沈萍. 框式动词及对外汉语教学[C]. 第六届东亚汉语教学研究生论坛暨第九届北京地区对外汉语教学研究生学术论坛论文集，2016.

附录

附表 1　语义角色引导词表

	语义角色	引导词
1	邻体	v 前引导词：和、跟、与、同、连、对、为、给、替、向、管、问、给、拿、朝 v 后引导词：于、给
2	工具	v 前引导词：用、以、拿
3	材料	v 前引导词：用、以、拿
4	方式	v 前引导词：用、以、通过
5	依据	v 前引导词：据、按、根据、按照、依、基于、照、依据、由、从、凭、随、趁、乘、借、靠
6	源点	v 前引导词：从、由、自、打、自从、打从、自打 v 后引导词：于、自
7	终点	v 前引导词：向、朝、往、奔、从、打、打从 v 后引导词：到、至

王诚文(1992—)，博士研究生，主要研究领域为句法语义分析、语言工程。

E-mail：chengwen_wang15@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

钱青青(1996—)，硕士研究生，主要研究领域为计算语言学。

E-mail：qianqingqing1996@foxmail.com

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net