现代汉语常用动词语义框架词典构建[1]

关同峰[1] 张坤丽[1] 段雪敏[1] 昝红英[1] 穗志方[2]

1 郑州大学信息工程学院，河南郑州 450001；2 北京大学计算语言学研究所，北京 100871

E-mail 215998315@qq.com ieklzhang@zzu.edu.cn xueminduan@stu.zzu.edu.cn iehyzan@zzu.edu.cn szf@pku.edu.cn

摘　要： 信息抽取、情感分析、机器翻译等自然语言处理任务的发展，离不开语义词典和语义框架的支持，因而构建出覆盖丰富语义知识的现代汉语常用动词语义框架词典显得尤为重要。本文在融合吸收已有研究成果的基础上，定义了动词语义框架词典结构，根据谓语论旨角色的不同，将语义框架分为基本语义框架和扩展语义框架；在大规模词义及论旨角色标注语料库的基础上，以语义为处理单元，自动抽取并归纳基本语义框架和扩展语义框架，辅以人工校对，构建了完整版和简化版动词语义框架词典。所构建的动词语义框架词典已出具规模，共包含 2782 个常用动词（4516 个义项）的语义框架详细描述及对应的例句。

关键词：现代汉语常用动词，语义框架词典，自动抽取，基本语义框架，扩展语义框架

Construction of the Dictionary of Commonly Used Verbal Semantic Frameworks in Modern Chinese

GUAN Tong-feng[1] ZHANG Kun-li[1] DUAN Xue-min[1] ZAN Hong-ying[1] SUI Zhi-fang[2]

1.School of Information Engineering of Zhengzhou University 2. Institute of Computational Linguistics of PEKING University

E-mail 215998315@qq.com ieklzhang@zzu.edu.cn xueminduan@stu.zzu.edu.cn iehyzan@zzu.edu.cn szf@pku.edu.cn

Abstract: The development of hot fields of natural language processing such as information extraction, sentiment analysis and machine translation are inseparable from the support of semantic lexicon and semantic framework. Therefore, it is especially important to construct the modern Chinese common verbs semantic framework dictionary covering rich semantic knowledge. Based on the fusion of existing research results, this paper defines the lexical framework of verbs. According to the different roles of predicate, the semantic framework is divided into basic semantic framework and extended semantic framework. On the basis of large-scale lexical and thematic role labeling corpus, the semantics is used as the processing unit, and the basic semantic framework and extended semantic framework are automatically extracted and summarized, supplemented by manual proofreading, and a complete and simplified version of the verb-word framework is constructed. The constructed lexical framework of the verb has been produced in a scale, which contains a detailed description of the semantic framework of 2782 commonly used verbs (4516 meanings) and corresponding example sentences.

Key words: modern Chinese common verbs, semantic framework dictionary, automatically extract, basic semantic framework, extended semantic framework

1 引言

自然语言处理（Natural Language Processing，NLP），旨在利用计算机分析自然语言语句和文本，抽取重要信息，进行分析、检索、阅读、问答、翻译和生成等。为了达成目标，学者们在词性标注、句法分析和语义依存分析等基础研究领域进行了深入研究，在信息抽取、情感分析、机器翻译、文本摘要、阅读理解和自动问答等应用研究领域也取得了长足的进步。

词汇语义知识库的主要成果有 WordNet[1]、FrameNet[2]和 MindNet[3]等；国内也拥有着众多的词典和知识库，包括《现代汉语语义词典》（the Chinese semantics Dictionary, CSD）[4]、《中文概念词典》（Chinese Concept Dictionary，CCD）[5]、现代汉语虚词用法知识库（the Chinese Function words Knowledge Base，CFKB）[6][7]和大规模词汇词义知识库（the Chinese Large-Scale Knowledge Base，CLSKB）[8]等。以动词

[1] 本文承 973 课题（2014CB340504）资助。

为主要研究对象的知识库有 Chinese FrameNet（CFN）、Mandarin VerbNet（MV）。CFN[9]是一种 FrameNet 风格的汉语框架语义网，参照 FrameNet，译建或创建适合汉语语义内容的框架，定义框架-框架关系，但是其定义的框架数目较少，仅有 309 个框架；MV 是由刘美君等[10]采用基于框架的构造方法来进行语言语义学研究的成果，完成了对“诱发运动”、“认知”、“沟通”、“情感”、“运动”、“知觉”、“社交联系”等主要动词类别的研究，证明了基于框架的汉语动词和动词类结构分析方法在语言上具有很好的动机和成效，但是由于采用人工构建法，导致包含的分类较少。目前，语义知识库的构建多数采用人工构建的方法，未能将词义放在一定的组合框架中去观察，往往采用静态的聚合分类法，加上为数不多的属性描述，使得构建规模和更新速度无法满足现在 NLP 发展的需求。

动词语义框架，是词典中动词与其论旨角色组合而成的语义关系结构系统。动词存在于每一个完整的句子之中，表示人或事物的动作、发展、变化、存在或者消亡的词。论旨角色分为核心论旨角色和外围论旨角色，必不可少的核心论旨角色，包含动作的发出者，以及动作的承受者，或者两者之一；而外围论旨角色则是为丰富句子语境，增加句子的画面感而补充的角色，包括工具、材料、方式、原因、目的、空间或时间上的起点和终点等。

以国家重点基础研究发展计划 973 课题(2014CB340504)为契机，在构建现代汉语广谱语义词典（the Broad-Spectrum semantic dictionary of Contemporary Chinese，BCSD）的背景下，本文融合吸收已有研究成果，定义了动词语义框架词典结构，根据谓语论旨角色的不同，将语义框架分为基本语义框架和扩展语义框架；以语义为处理单元，从大规模词义及论旨角色标注语料库中，自动抽取并归纳基本语义框架和扩展语义框架，辅以人工校对，构建了动词语义框架词典；在研究动词实现结构的差异时，完成了框架简化工作；构建出共包含 2782 个常用动词（4516 个义项）的语义框架详细描述及对应的例句。

2 动词语义框架

谓语动词在句子中起到凝聚作用，它可以把其它各种相关成分聚合在一起，句子的构建也必须围绕着动词这一核心，因此动词是句子的灵魂。同时动词对其论元的语义类存在约束，构建动词的基本语义框架，有利于把握人类对语言知识的认知模型。在构建动词语义词典时，充分吸收和借鉴现有的研究成果，并在此基础上对现有词类的语义属性描述框架进行扩充和完善。

动词语义框架的构建参照《现代汉语词典》第五版（XH5）的收词原则和词义划分颗粒度，一方面有机融合和继承现代《现代汉语语法信息词典》（the Grammar Knowledge Base，GKB）[11]、《现代汉语语义词典》和大规模词汇词义知识库的词类体系、语义分类体系及相关语义属性描述信息，以减少构建过程所需要的时间和人力，保证词典的高质量和高可信度。动词语义词典选定《现代汉语词典》第五版与《现代汉语语法信息词典》的交集动词中的 2782 个作为研究对象，其中包括了 1865 个单义词和 917 个多义词，依据《现代汉语谓词语义角色标注语料库规范》（以下简称 973 规范）中“谓词论旨角色层级分类体系”（图 1 所示）构建的一个动词与其论旨角色组合而成的语义关系结构系统。

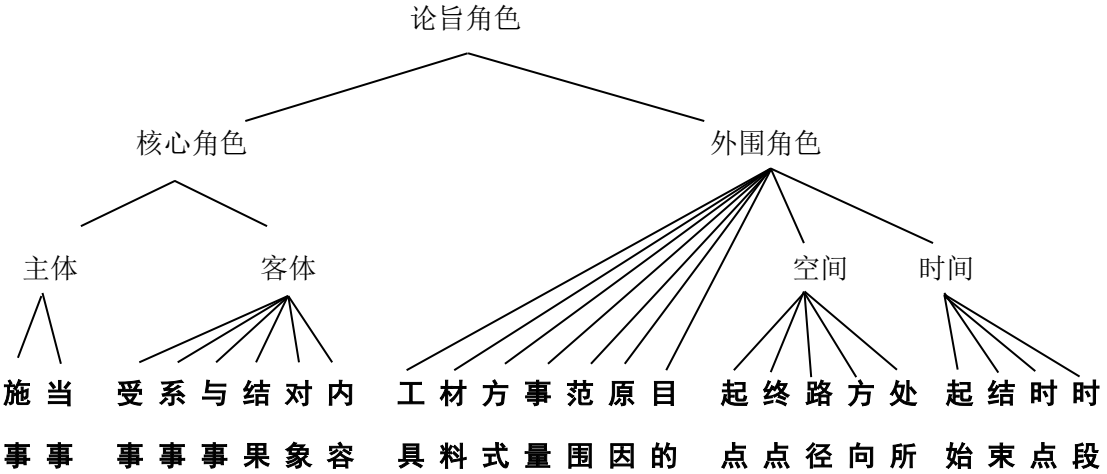

图 1 谓词论旨角色层级分类体系

论旨角色按照其与动词的紧密程度的差异，可以分为两部分：核心论旨角色和外围论旨角色。核心论旨角色，是与动词关系密切的必需成分，若缺少这部分成分，动词无法单独完整的叙述，核心论旨角色包含施事、当事、受事、系事、与事、结果、对象、内容；外围论旨角色，是与动词关系较疏远的成分，若缺少这部分成分，由动词与核心的论旨角色仍可完整叙述事件，只是内容上不全面，因此增加上这些成分之后，叙述会更加生动立体，外围论旨角色包含空间和时间这两个二级论旨角色和十六个三级论旨角色。

2.1 词典基本结构

动词库的基本架构如表 1 所示，主要包含动词条目、语义结构及论旨角色三部分，其中“动词条目”包含，动词词目及对词目的注音、释义、示例等内容，“语义结构”包含基本语义结构、扩展语义结构两类，而“论旨角色”含 8 个核心论旨角色、18 个外围论旨角色（总计 24 个三级论旨角色，2 个二级论旨角色）。

表 1 “动词库”语义结构基本结构

类别	条目	说明	样例			
动词条目	词目		定做			
	注音		dìngzuò			
	释义		专为某人或某事制作(物品)			
	示例		~生日蛋糕			
语义结构	基本语义结构	例句	Y76955:置于 [# 定做 #] 的 [%结果 楠木 盒 %] 中 ，			
		实现结构	[pred][结果]			
		典型结构	[pred][结果]			
	扩展语义结构	例句	Y31028:[%结果 那 簿子 %] 是 [%施事 外国 公司 %] [%目的 为 客户 %] [# 定做 #] 的 ，	Y18113:表示 将 [%工具 用 这 笔 捐款 %] [%目的 为 下岗 女工 %] 专门 [# 定做 #] [%结果 一 批 儿童 专用 接送车 %] ，	Y16348:可以 [%施事 由 大新 的 裁缝 师傅 %] [%目的 为 您 %] 度 身 [# 定做 #] ，	Y9872:从而 [%施事 她 %] 抛弃 了 [%目的 为 “ 社会 形象 ” %] 而 [# 定做 #] 的 [%结果 假 面 %] ，
		实现结构	[结果][施事][目的][pred]	[工具][目的][pred][结果]	[施事][目的][pred]	[施事][目的][pred][结果]
		典型结构	{[工具]}[施事]{[目的]}[pred][结果]			
论旨角色	核心论旨角色	施事	外国 公司		由 大新 的 裁缝 师傅	她
		当事				
		受事				
		结果	楠木 盒 | 那 簿子	一 批 儿童 专用 接送车		假 面
		……				
	外围论旨角色	工具		用 这 笔 捐款		
		材料				
		方式				
		事量				
		目的	为 客户	为 下岗 女工	为 您	为 “ 社会 形象 ”
		……				

2.2 结构描述

动词条目部分直接从《现代汉语词典》第五版中抽取而来。语义结构是指动词与其论旨角色组合而成的语义关系结构，含基本语义结构和扩展语义结构。动词语义框架词典需要为动词匹配相应的例句，然后从例句中自动抽取基本语义框架和扩展语义框架，因此词典需要将动词基本信息和相应的语义结构及对应的论旨角色分类整理，表中某些动词会因语料规模的不足，出现例句为空的现象。基本语义结构仅涉及包含核心论旨角色的语义关系结构，扩展语义结构则包含外围论旨角色。

3 语义框架抽取

无论是基本语义结构还是扩展语义结构，两者都有多种“实现结构框架”，因此需要选出一种典型的结构框架作为代表，动词语义框架词典要抽取的就是典型结构框架。

动词语义框架和语义角色标注语料库的构建是一个迭代过程：首先，针对标注语料库特点和框架词典规范，制定出动词框架和语义角色标注规范；其次，确定自动框架抽取方法；然后，每周讨论标注出现的问题，调整标注规范，用于指导接下来的语料库的标注，从而整理出新的动词框架。

语义框架从北京大学中文树库[12]及人民日报标注语料中抽取，其中北京大学中文树库的 55,764 句子中根据汉语语法点抽取 10,634 句进行语义和论旨角色标注作为语料（简称一万句），人民日报 2000 年 1 月份和 1998 年 1 月份抽取已标注目标动词语义及论旨角色的语料两批（简称为人民日报和 new），共计 42000 句。

在对语义框架的处理过程中，主要包含数据预处理、基本语义框架处理、扩展语义框架处理及简化版本四个过程，如图 2 所示。其中单义词可以看作仅有一个义项的多义词进行处理，在处理中，若出现多个例句排序，默认按“一万句->人民日报->new”优先顺序选择典型例句。

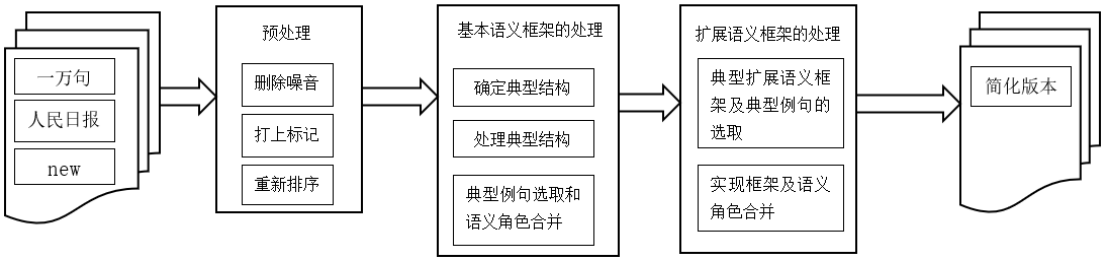

图 2 框架处理过程

3.1 数据预处理

数据预处理工作主要步骤为删除无用例句、标记边界、重排序、过滤噪声。重排序，是将实现框架中相同的排列在一起。过滤噪声，是在计算典型框架之前，先按框架的合理性进行过滤，在 973 规范中，对主体和客体的对应关系作了一定的约束，如表 2 所示。处理之后，单义词仅保留三个以上例句的词语，多义词删除无例句的义项。

表 2 主体、客体对应关系表

客体	主体	对应关系说明
系事	当事	系事不能与施事对应
受事	施事	不能与当事对应
与事	施事	通常不与当事对应
结果	施事|当事	可与施事和当事对应
内容	施事|当事	可与施事和当事对应
对象	施事|当事	可与施事和当事对应

3.2 基本语义框架处理

论旨角色在实际语句中所处的位置较为灵活，因而一个动词的某一个义项可以有多个由于论旨角色所处位置不同而形成的排列顺序不同的语义结构，动词库将其视为“实现的语义结构”。由“实现的语义

结构”抽象出的语义结构视为“典型的语义结构”。“实现的语义结构”与“典型的语义结构”是语义结构的“一般变体”和“典型变体”，“典型的语义结构”按照论旨角色出现的逻辑顺序排列。

通俗来讲，实现结构就是组员，典型结构就是从组员中选取的班长，有了典型结构就可以方便认识该义项。基本语义框架的处理分成三个步骤，第一步是将所有的单义词按计算的方式选择出典型结构，第二步针对已经人工选择典型结构，则用人工选择的典型结构代替第一步中自动选择出的典型结构，第三步则针对第一步和第二步的结果确定相应实现结构的例句以及合并语义角色。

3.3 扩展语义框架处理

扩展语义框架由于包含 18 个外围角色，故会存在多个实现的语义结构，需要从其中抽取出典型语义结构。典型扩展语义框架原则上是在典型基本语义框架基础上增加相应的外围角色，但根据实际情况也作了一些调整，具体如图 3 所示。

图中①表示对典型基本语义框架是否是某些实现结构子集的判断，存在子集时，根据是否有多个实现结构来决定合并，分别选择②合并和③无需合并；交集在合并时还需要考虑核心角色是否相同的情况，不相同时应考虑④选择最长的实现结构同时例句数也是最多的框架。最终所有的动词可经过②③④到达结束状态。

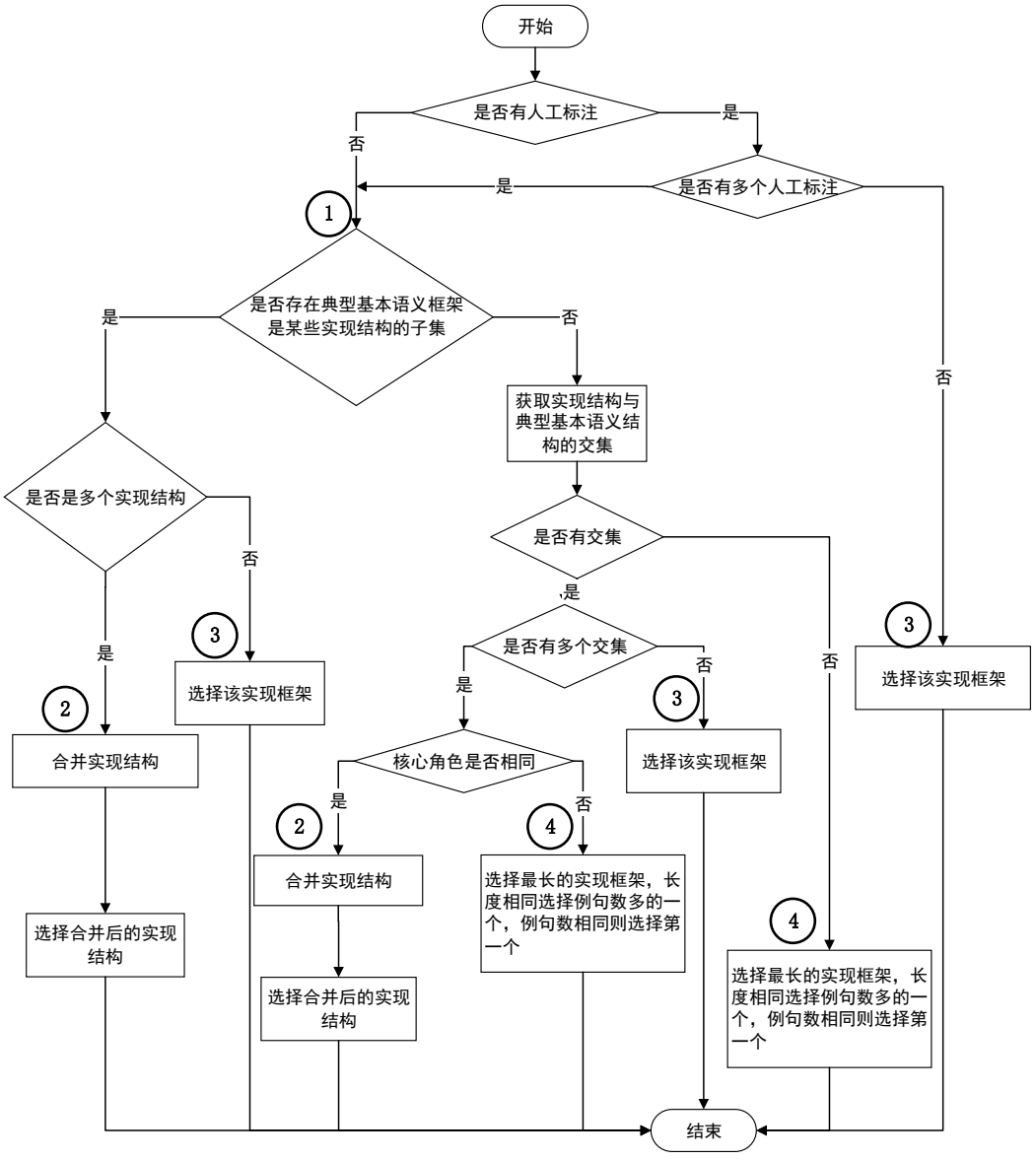

图 3 扩展语义框架处理流程图

3.4 框架简化

在完成基本语义框架和扩展语义框架的自动抽取工作之后，为了对比研究动词在不同句子中其实现结构的差异，探寻语义结构的内在规律，将抽取语义结构的结果进行简化。简化后的基本语义结构和扩展语义结构如表 3 表 4 所示。将实现结构合并在一个单元格中，框架之间用“|”分隔开，仅保留典型例句，同时将语义角色合并，合并的过程中去重，也即在同一个论旨角色中，不出现重复的内容，最终一个动词仅保留两行记录。

表 3“定做”“布阵”简化后的基本语义结构

动词	基本语义结构		
	例句	实现结构	典型结构
定做	Y76955:置于 [# 定做 #] 的 [%结果 楠木盒 %] 中 ，	[pred][结果]	[pred][结果]
布阵	N594136:[%施事 由 钱 三强 %] 点将 [# 布阵 #] 的 名单 还 很 长 ，	[当事][pred][施事][pred]	[施事][pred]

表 4“定做”“上升”简化后的扩展语义结构

动词	扩展语义结构		
	例句	实现结构	典型结构
定做	Y18113:表示 将 [%工具 用 这 笔 捐款 %] [%目的 为 下岗 女工 %] 专门 [# 定做 #] [%结果 一 批 儿童 专用 接送车 %] ，	[结果][施事][目的][pred][工具][目的][pred][结果][施事][目的][pred][施事][目的][pred][结果]	{[工具]}[施事]{[目的]}[pred][结果]
上升	N606207: 然而 笔者 认为 [%当事 这 一 斗争 %] 或许 并不 能 [# 上升 #] [%终点 到 集体 的 爱国主义 的 高度 %] 。	[当事][pred][终点][时段][事量][pred][起始][当事][pred][时间][当事][pred]	{[起始][时间]}[当事][pred]{[终点]}

4 结果及分析

对单义词和多义词抽取典型框架的结构进行统计分析，统计结构如表 5 所示。表中的义项总数即每类动词所包含的所有义项数目，因每个多义词包含多个义项，故 917 个多义词总共有 2651 个义项。

在这些义项中，同时包含基本语义框架和扩展语义框架的有 785 个，仅包含基本语义框架的由 672 个，仅包含扩展语义框架的有 208 个，还存在 986 个义项两者皆无的情况，抽取这部分义项进行分析（表 6 所示），发现基本上由两种原因造成，第一个原因是部分动词以单字形式出现，比如说上、下、交、令、休、会、做等，这部分动词的某些义项已经不常使用或者使用过少，因而未在语料中出现；也存在包含这些字的动词，比如说上台、上学、下船、下岗、交换、交替等，分别包含上、下、交的意思，因而在对语料进行分词时，将其作为整体处理。第二个原因是使用的语料库以新闻为主，这些词在那段时间的新闻中不出现，比如说上升、上岗、临门、修养等，之后的工作考虑增加语料库的种类，丰富例句数。

表 5 语义框架抽取结果统计

类别	词语数	筛选处理	义项总数	仅有基本语义框架	仅有扩展语义框架	两者都有	两者都无
多义词	917	-	2651	672	208	785	986
单义词	1865	190（例句数低于 3）	1865	236	37	1592	0
合计	2782	190	4516	908	245	2377	986

表 6 基本语义框架和扩展语义框架都没有的义项

词目	释义	示例	词目	释义	示例
上	到；去(某个地方)	~街|~工厂|他~哪儿去了	上升	由低处往高处移动	一缕炊烟袅袅~
上	向上级呈递	~书	上岗	到执行守卫、警戒等任务的岗位	警察~指挥交通
下	由高处到低处	~山|~楼|顺流而~	临门	到达球门前	~一脚
下	发布；投递	~命令|~通知|~战书	主持	主张；维护	~公道|~正义
交	到(某一时辰或季节)	~子时|明天就~冬至了|~九的天气	了解	打听；调查	先去~情况
令	命令	~各校切实执行	休养	恢复并发展国家或人民的经济力量	~民力

5 结论

本文定义了动词语义框架词典结构，将语义框架分为基本语义框架和扩展语义框架，同时在大规模词义及论旨角色标注语料库的基础上，自动抽取并归纳出基本语义框架和扩展语义框架，构建出了包含917 个多义词和 1865 个单义词的现代汉语常用动词语义框架词典，为自然语言处理的上层任务如机器翻译、情感分析和信息抽取等提供帮助。通过分析已构建的动词语义框架词典，发现多义词的某些义项，存在既未包含基本语义框架，也未包含扩展语义框架的情况，主要是由于语料库的来源为新闻，下一步工作中，考虑增加其他数据源的语料，进而扩充例句的多样性，从多样化的实现框架中抽取更具有代表性的典型框架，以提高动词语义框架的质量；同时，扩展动词语义框架词典的规模，期望覆盖更多数量的动词条目。

参 考 文 献

[1] Fellbaum C. Wordnet: An electronic lexical database [M]. Cambridge, Mass.: MIT Press,1999.

[2] Bake, C.F, C.J. Fillmore, and John B.Lowe. The Berkeley FrameNet Project, In Proceedings of COLING'98, 1998, pp.86-90

[3] Richardson S D, Dolan W B, Vanderwende L. MindNet: acquiring and structuring semantic information from text[C]// ACL. 1998:1098-1102.

[4] 王惠,詹卫东, 俞士汶. 现代汉语语义词典规范[J]. 汉语语言与计算学报，2003，13(2):159-176

[5] 刘扬, 俞士汶, 于江生. CCD 语义知识库的构造研究[J]. 小型微型计算机系统, 2005, 26(8):1411-1415.

[6] Zan H Y, Zhang K L, Zhu X F, et al. Research on the Chinese Function Word Usage Knowledge Base[J]. International Journal on Asian Language Processing, 2011, 21（4）:185-198.

[7] 张坤丽,昝红英，柴玉梅，韩英杰，赵丹.现代汉语虚词用法知识库建设综述[J].中文信息学报,2015,29(3):1-8.

[8] 石金铭, 昝红英, 韩英杰. 大规模汉语词汇语义知识库的构建[J]. 山西大学学报自然科学版, 2015, 38(4):581-587.

[9] You LP, Liu KY. Building Chinese FrameNet Database[C]. In:Proceedings of 2005 IEEE NLP-KE, 2005:301-306.

[10] Liu, Meichun. Forthcoming. Semantic Annotation and the Mandarin VerbNet. To appear in Chinese Language Resources: Data Collection, Linguistic Analysis, Annotation and Language Processing. Chu-Ren Huang, Shu-Kai Hsieh, and Peng Jin (eds.). Springer.

[11] 俞士汶,朱学锋,王慧,等.现代汉语语法信息词典详解[M].北京：清华大学出版社，2003.

[12] Zhan W D. Peking University Treebank[M]/Encyclopedia of Chinese Language and Linguistics, Volume 3. General Editor Rint Sybesma. Netherlands: Brill Publishing House,2016:332-336.