

文章编号: 1003-0077(2023)01-0016-17

一种改进的汉语语义角色分类体系与标注实践

宋 衡^{1,2}, 曹存根¹, 王 亚^{1,2}, 王 石¹

(1.中国科学院 计算技术研究所 智能信息处理重点实验室,北京 100190;

2.中国科学院大学,北京 100049)

摘 要: 语义角色标注是自然语言处理中的一项重要任务,涉及文本挖掘、神经网络学习、统计语言学等多个研究领域。在对一个汉语语料库进行语义角色分析时,发现现有的汉语语义角色体系存在一定的局限性。通过引入新的语义角色、吸收和重新定义现有语义角色体系中的语义角色,该文提出了一种改进的汉语语义角色分类体系。该语义角色分类体系将事件中的语义角色分为两大类,即中枢语义角色和周边语义角色,其中,周边语义角色可进一步被细分为主要周边语义角色和辅助周边语义角色。为了减少语义理解的主观性从而客观地判断语义角色类型,该文基于语义和句式以“判断标准—相应例句”的形式详细解释了语义角色分类体系中的主要周边语义角色,并从中枢语义角色半自动化判断、复合动词作中枢语义角色处理、易混淆语义角色难点分析和辅助周边语义角色标注规定等多个方面给出了提高语义角色标注一致性的说明。最后,根据提出的语义角色分类体系对实验语料库进行了语义角色的标注实践,分析了标注一致性,统计了新提出和重定义的主要周边语义角色在语料库中的分布情况,并与基于鲁川语义角色分类体系得到的标注结果进行了对比。

关键词: 汉语语义角色分类体系;知识抽取;文本挖掘;浅层语义分析;语料库

中图分类号: TP391

文献标识码: A

A Fine-Grained Annotated Dataset for Chinese Semantic-Role Labeling

SONG Heng^{1,2}, CAO Cungen¹, WANG Ya^{1,2}, WANG Shi¹

(1. Key Laboratory of Intelligent Information Processing, Institute of Computer Technology,

Chinese Academy of Sciences, Beijing 100190, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Semantic roles play an important role in the natural language understanding, but most of the existing semantic-role training datasets are relatively rough or even misleading in labeling semantic roles. In order to facilitate the fine-grained semantic analysis, an improved taxonomy of Chinese semantic roles is proposed by investigating a real-world corpus. Focusing on a corpus formed with sentences with only one pivotal semantic role, we propose a semi-automatic method for fine-grained Chinese semantic role dataset construction. A corpus of 9550 sentences has been labeled with 9423 pivot semantic roles, 29142 principal peripheral semantic roles and 3745 auxiliary peripheral semantic roles. Among them, 172 sentences are double-labeled with semantic roles and 104 sentences are labeled with semantic roles of uncertain semantic events. With a Bi-LSTM+CRF model, we compare the dataset against the Chinese Proposition Bank and reveal differences in the recognition of principal peripheral semantic roles, which provide clues for further improvement.

Keywords: Chinese semantic roles taxonomy; knowledge extraction; text mining; shallow semantic analysis; corpus

收稿日期: 2021-07-05 定稿日期: 2021-07-29

基金项目: 国家重点研发计划(2017YFC1700302, 2017YFB1002300);国家自然科学基金(61702234);北京市科技新星计划交叉学科合作课题(Z191100001119014)

0 引言

语义分析是自然语言处理领域的重要任务,它致力于获取给定句子蕴含的语义信息,并以计算机能理解的方式进行表示^[1]。目前,语义分析的研究工作主要包括深层语义分析和浅层语义分析。深层语义分析的主要工作有普适概念认知标注(Universal Conceptual Cognitive Annotation, UCCA)、语义依存分析(Semantic Dependency Parsing, SDP)和抽象语义表示(Abstract Meaning Representation, ARM)等。然而,深层语义分析存在语义层次涉及范围广、歧义性和模糊性大、难以用良好形式化方法刻画所有语义信息等问题,甚至受限于当前技术水平,短期内不易形成具有较强实用性的成果^[2]。在浅层语义分析方面,目前主要的实现方式是语义角色标注(Semantic Role Labeling, SRL)^[3],该方法具有多语言通用和表现形式自然等诸多优点。

作为实现浅层语义分析的基础,定义一个具体且清晰的语义角色分类体系来解释句子成分的语义显得至关重要,它能够促进语义角色数据集的建立以及相关语义分析模型的研究。为了适应汉语语义分析的需要,中国学者已经提出了多种语义角色分类体系,具有代表性的工作有朱晓亚的《现代汉语句模研究》^[4]、袁毓林的《基于认知的汉语计算语言学研究》^[5]、鲁川的《知识工程语言学》^[6]以及刘茂福等人的《基于认知与计算的事件语义学研究》^[7]等。针对 Web 文本语义分析和知识抽取任务中的实际需求^[8-12],中科院计算所曹存根团队利用上述语义角色分类体系构建了一种基于语义分类和描述框架(Framework of Semantic Taxonomy and Description, FSTD)的知识库^[8-9]。FSTD 设计了一种不限于具体领域的通用语义文法。该文法是一组由语义非终结符、词类非终结符和常量终结符组成的序列。其中,语义非终结符是形为“<语义角色: 类型>”的非终结符,“类型”充当语义角色对象的类型;词类非终结符表示一组语义相同的词语;常量终结符是一组字符串,可以是一个或多个词语。例如,<施事: 认知主体><买词类>[了]<受事: 物体|服务>表示 FSTD 中“买”事件的一个语义文法。其中,“<施事: 认知主体>”和“<受事: 物体|服务>”为文法中的语义非终结符;“<买词类>”为词类非终结符;“[了]”为常量终结符。施事后的对象类型“认知主体”和受事后的对象类型“物体|服务”为对语义角色

类型所做的语义限制。语义角色是语义层面上的定义,而对象类型则是知识层面上的定义。利用 FSTD,我们进行了相关科研工作,如 Zang 等^[8]和方芳^[10]的知识抽取、曹聪^[11]和王亚等人^[12]的常识知识获取等。

FSTD 语义文法是以鲁川语义角色分类体系为主,并参考其他体系而设计的。然而,随着研究的深入,FSTD 语义文法所使用的语义角色分类体系逐渐无法满足 Web 文本语义分析与知识抽取的需求^[10-11],因此,需要对汉语语义角色的分类体系进行进一步研究。曹存根团队多年来的科研工作^[8, 10-12]逐渐积累了一个良好的中文语义语料库(Chinese Semantic Corpus, CSC),其语料来源于商业网站、门户网站和新闻网站等,涉及汽车^①和数码产品^②的介绍、各种新闻^③以及百科知识^④文档等,共含有 9 541 条汉语句。基于现有的语义角色分类体系^[4-7],本文对 CSC 进行了详细的分析和调研,发现如下四类问题。

首先,缺乏统一的语义角色分类体系。语义理解的主观性导致不同体系在语义角色的分类数量和术语内涵限定上存在较大差异。例如,文献^[4-7]中分别定义了 14、17、16 和 26 种周边语义角色。此外,使用不同语义角色体系分析 CSC 中的语料时,常会出现同一句子成分在不同体系中名称不同的问题。例如,对于句子“水果都腐烂了”,“水果”在鲁川语义角色体系中是当事,而在袁毓林语义角色体系中是主事。本文综合现有语义角色体系的优点,提出了一种改进的汉语语义角色分类体系,并从语义和判断标准的角度说明分类的原因,例如现有文献^[5-7]将时间、空间和发展过程的归宿点理解为语义角色宿事,但在对 CSC 进行分析过程中,我们发现主体加入的组织和活动具有类似的归宿点语义信息,因此可以将主体参加的组织和活动归类到宿事。本文将在第 3 节详述这个问题。

其次,语义角色标注句子成分的方式比较单一。使用现有语义角色分类体系进行标注时,仅为句子的每个成分标注一个语义角色,而在实际情况中,某些句子成分可能同时扮演多个语义角色。例如,对于句子“武警击毙了抢劫犯”和“武警误杀了人质”,从常规语义角色标注的角度来看,两个句子中的成

① 汽车之家, <https://www.autohome.com.cn/>

② 泡泡网, <http://www.pcpop.com/>

③ 新浪新闻, <https://news.sina.com.cn/>

④ 百度百科, <https://baike.baidu.com/>

分“武警”都是施事,但从语义角色受损还是受益的角度来看,前一个句子中的“武警”是受益者,而后一个句子中的“武警”是受损者。为此,本文提出辅助周边语义角色的概念,并将在第 2 节说明这个问题。

再次,部分定义模糊的语义角色使得句中某些成分难以被准确标注,并导致部分事件的 FSTD 语义文法被错误编写。事实上,通过对 CSC 进行分析,本文发现该语料库对语义角色“客事”的标注错误率高达 17.0%,对语义角色“受事”的标注错误率为 14.86%。为了提高 FSTD 语义文法编写和语义角色标注的准确性和一致性,本文明确给出了语义角色的判断标准及其对应例句。“判断标准-相应例句”表达方式的提出,有助于减少语义理解主观性,帮助标注人员尽可能以客观的角度进行语义角色标注。此外,本文还从中枢语义角色的半自动化判断、复合动词作中枢语义角色的处理、易混淆主要周边语义角色难点分析和辅助周边语义角色标注规定等四个方面进行了提高语义角色标注一致性的说明,本文在第 3、4 节将详述这个问题。

最后,语义角色种类不够丰富将导致标注的语料存在语义偏差。通过对 CSC 进行分析,本文发现现有语义角色分类体系无法解释一些特殊句子,解决此问题需要引入新的语义角色。先前的工作^[10-11]也表明,基于现有体系所创建的 FSTD 语义文法无法满足 Web 文本语义分析与知识抽取的实际需求。例如,对于句子“我叔叔建了一栋别墅”,“我叔叔”不是真正的施事,因为他雇佣了一个建筑队来建别墅。为此,本文引入了一种新的语义角色,称为雇施事;又如,在句子“计算机病毒窃取了我们的敏感信息”中,“计算机病毒”不是真正的施事,因此提出了一种被称为代施事的语义角色。本文将在第 3 节详述该问题。

本文组织结构如下。第 1 节介绍了相关工作。第 2 节概述事件中语义角色的上层分类。第 3 节重点讨论主要周边语义角色的具体分类,这是语义角色分类体系中最重要分支。第 4 节进行了提高语义角色标注一致性的说明。第 5 节进行了语义角色标注实践与分析。第 6 节总结了本文工作并对未来的研究进行了展望。

1 相关研究

1.1 国外相关工作

1968 年,Quillian 从心理学的角度提出了语义

网络理论^[13]。语义网络是一种以网络格式表示人们知识的结构化模型。在语义网络中,知识由相互连接的节点和边表示,其中节点表示对象和概念,边表示节点之间的关系。为了从语义角度弥补转换成语法的不足,Fillmore 提出了一种能够表示句法结构和语义间关系的语言学理论,即著名的格文法(Case Grammar, CG)^[14]。Fillmore 认为“格”能够真正构成自然语言深层结构中的语法关系,其主要包括施事格、工具格、与格、客体格、方位格、结果格等,“格”本质上就是语义角色。格文法将动词作为中心,每个动词可支配一定的“格”。

语义网络理论和格文法为自然语言语义的研究与应用奠定了坚实的基础,其中比较著名的语义研究资源有 WordNet^[15-16]、FrameNet^[17-18]以及 VerbNet^[19-20]。目前,这三种语义型资源已经在知识表示等自然语言处理领域被广泛应用,表 1 展示了它们在语义特点方面的差异。

表 1 WordNet、FrameNet 以及 VerbNet 对比

特点	WordNet	FrameNet	VerbNet
研究对象	单词	框架	句法
句法	无	有	有
语义关系	同义、反义、整体部分、蕴含、上下位	使用、继承、总分、透视、优先、参见及使动	上下位
语义表达	比较浅层	深层	深层
上下文表达语义	无	有	有
应用领域	词义标注、文本校对、知识推理	语义分析、机器翻译、机器问答、信息抽取	语义角色标注、动词自动抽取、语义分析

1.2 国内相关工作

国内学者针对汉语的特点对语义网络理论和格文法进行了大量研究。鲁川改进了语义网络和格文法的内容和形式,提出了意合网络的概念^[21]。意合网络是一种由语义角色单位组成,能够准确表示语义关系的层次网络。作为汉语句子的基本描述单元,语义角色在文本语义分析和知识抽取中起着至关重要的作用。目前,国内研究者已经建立了几种不同的汉语语义角色分类体系。

朱晓亚^[4]认为事件的语义结构由动核及其相关的动词元语义成分构成。她提出从动词的语义范畴

来确定动词在动词核心结构中的语义角色,并通过动词核心语义角色间的搭配建立基本的语义结构类型。朱晓亚共定义了 14 种语义角色。

基于认知的角度以及计算机处理汉语的实际需求,袁毓林^[5]总结了现代汉语的 17 种语义角色类型,并详述了这些语义角色广泛的语义定义和突出的句法特征。此外,袁毓林还使用多种动态语义和语法特征来区分所定义的语义角色。

鲁川^[6]认为事件是知识的基本单位,而事件由一个语义角色为中枢,一个或多个相关语义角色为周边所构成。鲁川将中枢语义角色分为 26 类,并将其作为现代汉语基本句模的分类标准。基于穷尽性、系统性、实用性和顺序性的分类原则,鲁川总结了 26 种周边语义角色。

刘茂福等^[7]认为原子事件是事件语义分析计算的基本信息处理单位。在认知科学和题元理论的基础上,刘茂福等讨论了原子事件的语义角色,并总结了 16 种原子事件的语义角色。相比于袁毓林^[5]语义角色分类体系,刘茂福等删除了感事这个语义角色,将一部分感事归入施事,另一部分归入为主事。

在综合参考不同汉语语义角色体系^[5-6]的基础上,王丽杰^[22]根据自身需求对现有汉语语义体系进行了语义角色的补充,总结了 35 种语义角色类型。值得注意的是,这些语义角色不仅包括事件中的周边语义角色,还包括事件与事件之间的关系,如并列、让步和转折等。本文主要研究的是事件中语义角色及其分类,因此对事件与事件间的关系不做过多讨论。

为解决中文命题库语义角色不一致的问题,柏晓鹏^[23]在袁毓林^[5]的基础上,定义了一套全局性语义角色分类体系,并总结了 23 个全局性语义角色。然而,柏晓鹏定义的语义角色主要是为中文命题库服务,并不适用于所有情况^[23]。

基于对立互补的原则,王诚文等人^[24]以鲁川^[6]的语义角色体系为基础,并融合袁毓林^[5]体系,形成了新的融合语义角色分类体系。为了适应具体应用场景对外围语义角色精细度要求高的需求,王诚文的新融合体系包括 15 个语义角色,对核心语义角色分类较为宽泛,而对外围语义角色分类详细且覆盖面广。

2 事件中语义角色的上层分类

首先,在事件中语义角色分类的上层方面,本文

将事件的语义角色分为两大类:中枢语义角色和周边语义角色,如图 1 所示。

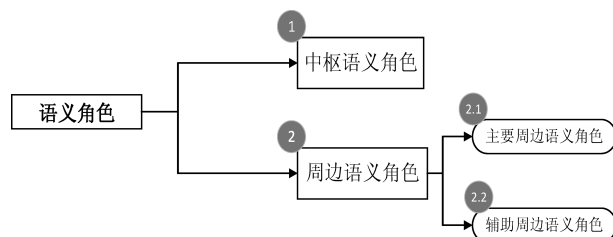
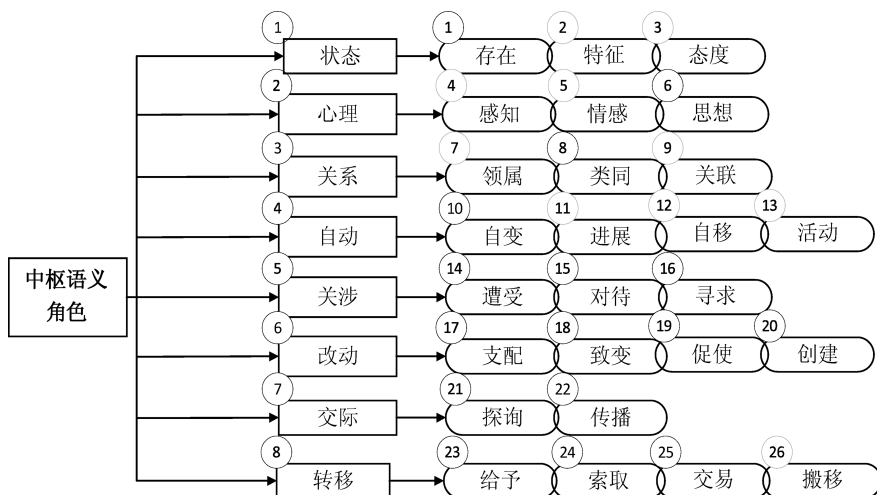


图 1 语义角色的上层分类

事件的中枢语义角色(Pivotal Semantic Role, PSR)是事件的中枢,它通常由谓词扮演。中枢语义角色是一个事件的主干,且一个事件只能有一个中枢语义角色。中枢语义角色在不同语义角色分类体系中具有不同的名称,例如,它在文献[4]中被称为动核,在文献[5]中被称为超级论元,在文献[6]中被称为中枢事元,在文献[7]中被称为事件谓词。考虑到 FSTD 在构建时使用了鲁川^[6]和王亚^[9]的中枢事元分类体系,并在实际 Web 文本语义分析和知识抽取任务^[8,10-12]应用中取得了良好的效果,因此,本文的汉语语义角色分类体系直接使用了他们的中枢事元分类体系,如图 2 所示,他们将中枢语义角色分为 26 种类别,限于篇幅,中枢语义角色的具体定义在本文不进行详细阐述。

周边语义角色是本文的研究重点,与中枢语义角色不同,它能够丰富和完善事件的语义结构和内涵。通过对 CSC 进行分析,本文发现现有汉语语义角色体系^[4-7]存在一定局限。例如,受益者和受损者是两种对语义理解有重要作用的语义角色,并且在语料中经常出现,然而,现有体系^[4-5,7]并没有考虑这两个语义角色;在文献[6]中,鲁川将受益者和受损者统一作为周边语义角色涉事处理,这种方式导致在语义角色标注实践中,不仅会出现无法区分语义角色是受益者还是受损者的问题,还会出现周边语义角色冲突的问题。例如,在句子“日军偷袭了珍珠港”中,珍珠港是受事,同时按照鲁川体系^[6]对涉事的定义,珍珠港还可以被标注为涉事,因为它被偷袭,表明其为事件中的受损者,这造成了珍珠港能够同时被标注为受事和涉事两种不同的周边语义角色。显然,这并不合理。为了解决该问题,本文将周边语义角色分为两类,即主要周边语义角色(Principal Peripheral Semantic Roles, PPSR)和辅助周边语义角色(Auxiliary Peripheral Semantic Roles, APSR),如图 1 所示。

图2 中枢语义角色的分类^[6,9]

主要周边语义角色 (PPSR)表示事件中中枢所关涉的处于周边的关键语义角色。一个事件可以存在多个主要周边语义角色。本文第3节将详述主要周边语义角色的分类体系。

辅助周边语义角色 (APSR)表示在事件中获得收益或者损失利益的语义角色,即受益者(Benefactive)或者受损者(Malefactive),如表2所示。事实上,在对CSC分析过程中,经常遇到一些事件角色在事件中损失利益或者获得收益。对于这些事件角色,除了需要标注必要的主要周边语义角色外,还应该为其标注相关的辅助周边语义角色。通过将事件中的主要周边语义角色和辅助周边语义角色相结合,可以掌握事件中更多的语义信息。

表2 辅助周边语义角色

角色	定义	例句
受益者	在事件中获得收益的语义角色	胡小姐祝我们新年快乐。 康熙皇帝特赦了罪犯。 我父亲给我买了一件外套。
受损者	在事件中损失利益的语义角色	邪教伤害了无数的家庭。 计算机病毒盗取了用户敏感信息。 日军偷袭了珍珠港。

3 主要周边语义角色新探

本节将讨论图1中语义角色分类体系中最复杂的分支,即主要周边语义角色。

3.1 主要周边语义角色的分类

在主要周边语义角色的一级子类方面,鲁川^[6]将

其分成8种类别,分别是主体(Nominative)、客体(Accusative)、邻体(Neighbor)、伴体(Accompany)、系体(Relative)、情由(Causative)、时空(Circumstance)和度量(Measure)。在这8种一级子类的基础上,本文引入了一个新的主要周边语义角色一级子类,即描述(Descriptive),旨在处理包含描述对象属性(Attributes)或者性质(Properties)的句子。引入描述这一新的周边语义角色一级子类,有利于促进信息检索、知识抽取、文本挖掘和常识获取等应用^[25]。

主要周边语义角色的一级子类可被进一步划分为32个更具体的二级子类,如图3所示。在这32个主要周边语义角色中,有19个主要周边语义角色(图3中以白色圆圈标记)已经在现有语义角色分类体系^[4-7]中被提出和定义,本文并未对其进行修改,但以“判断标准-相应例句”的形式进行了总结,如表3所示。“判断标准-相应例句”的总结方式有助于减少语义理解的主观性,从而提高语义角色的标注一致性。此外,本文提出了6个新的主要周边语义角色(图3中以黑色圆圈标记),即雇施事(Hiring Agent)、代施事(Resulting Agent)、变事(Change)、属性(Attribute)、性质(Property)和值事(Attribute-value),并且重新定义了7个现有的主要周边语义角色(图3中以灰蓝色圆圈标记),即受事(Patient)、客事(Entity)、源事(Origin)、宿事(Destination)、向事(Target)、用事(Reliance)和位事(Location)。本文将在第3.2节对它们进行详细解释和说明。

需要说明的是,本文语义角色分类体系中的每个“判断标准”在本质上都可作为一个更细粒度的语

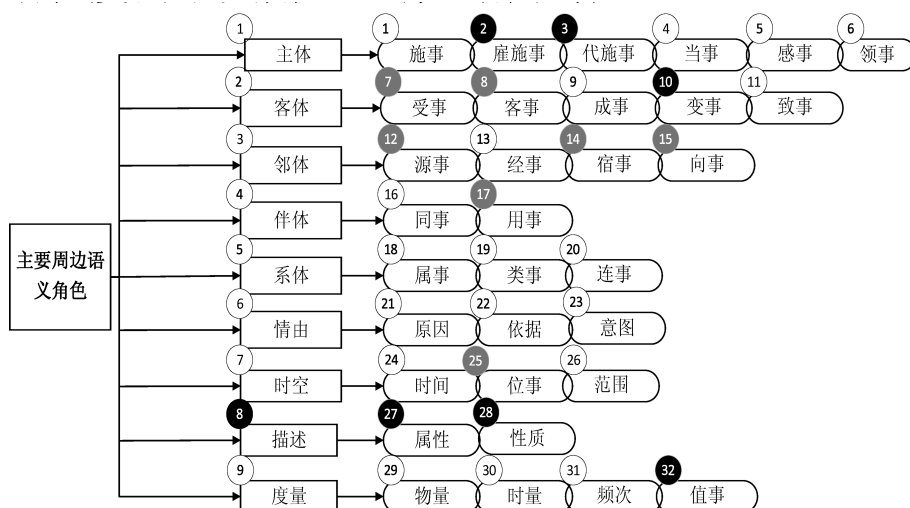


图3 本文提出的主要周边语义角色分类体系

义角色理解和标注,然而,过于细粒度的语义角色划分会增加语言资源构建者的负担,不利于大规模语言资源和知识库的快速构建,因此在实际应用中,可

根据具体应用场景和需求,灵活地选择是否需要提炼出相应的判断标准作为更细粒度的语义角色进行理解和标注。

表3 已有工作^[4-7]中的主要周边语义角色^①

主要周边语义角色	来源	判断标准-相应例句
施事 ^[4-7]	朱袁鲁刘	发出直接可控行为或思维活动的有意识的主体: ①自身亲自执行可控行动的有意识的主体(学生们飞快地奔向食堂); ②发出指令指使客体完成而自身并未亲自执行该指令的主体(唐永清逼迫小黑抢劫拍卖会); ③自身在进行思维活动的有意识的主体(王主任认真思考问题)
当事 ^[4-7]	朱袁鲁刘	发出非可控行为的主体或各种关系的主体: ①发出非可控行为的主体(黄河向东流去); ②表现自身状态的主体(计算所大楼好高); ③除领属关系 ^[6,9] 之外,各种关联关系的主体(计算所靠近融科资讯中心)
感事 ^[4-6]	朱袁鲁	非可控心理活动的主体: ①各种感知体验的主体(学生们听到了鸣炮声音); ②各种情感状态的主体(他热爱美丽的大自然)
领事 ^[5-7]	袁鲁刘	领属关系 ^[6,9] 的主体: ①领有关系的主体(李老太拥有一套北京四合院); ②整体部分关系的主体(小米手机搭载高通骁龙处理器)
成事 ^[4-7]	朱袁鲁刘	主体创建或产生的新客体以及主体实施参与事件产生的结果: ①有意识主体创建的新客体(联想公司制造了一款新电脑); ②有意识主体对客体进行加工产生的另一种状态的新客体(面点师傅把面团揉成了面条); ③有意识主体实施或参与事件产生的结果(弟弟期末考了90分); ④无意识主体产生的新客体(蚕的脑神经分泌一种脑激素)
致事 ^[4,6]	朱鲁	主体通过促使类行为 ^[6,9] 为实现自己意图所促使的客体,即主体为实现自己的意图,通过促使类行为导致客体去执行后继的行动(连长命令排长查岗;朝鲜请求中国派兵援助;作者邀请他为这本新书拟写书评)
经事 ^[4-7]	朱袁鲁刘	事件中经历的空间或过程: ①事件所经过的空间(该公司的许多欧洲航班都途经上海机场); ②事件所经历的过程(小明经过高考考上了大学)
同事 ^[4-7]	朱袁鲁刘	事件中的伴随者、排除者或作为基准的邻体 ^② : ①事件中的伴随者(我陪妈妈去买衣服); ②事件中的排除者(除了英语我还会日语); ③主体比较或测量的基准(姐姐比妹妹大五岁)

① 朱晓亚体系中的系事、袁毓林体系中的主事以及鲁川体系中的当事所指内涵一致,但术语命名相差较大,其他主要周边语义角色也有类似的情况,本表中主要周边语义角色术语的命名以鲁川体系为主。

② 现有某些语义角色体系又将作为基准的邻体作为新语义角色基准或者比事处理,此处考虑主要周边语义角色种类、数量方面的权衡,因此尊重鲁川语义角色体系,将作为基准的邻体归纳到同事。

续表

主要周边语义角色	来源	判断标准-相应例句
属事 ^[5-7]	袁鲁刘	领属关系 ^[6,9] 中所属的①亲属朋友(李阿姨有两个儿子);②敌人(老鼠有很多天敌);③财务(李老太拥有一套北京四合院);④构成部分(小米手机搭载高通骁龙处理器)
类事 ^[4-7]	朱袁鲁刘	事件中存在的类同者、类似者以及事件中主体最后变成的状态:①事件中客观存在或公认类同关系 ^[6,9] 的类同者(玫瑰是蔷薇科植物);②事件中的类似者(白鹤岛酷似一条硕大的鳊鱼);③有意志主体扮成的类同关系的类同者(警方缉毒组伪装成黑帮分子);④有意志主体主观认同的类同者(安禄山在洛阳自封大燕皇帝);⑤与主体在数值方面相当的客体(江苏省面积相当于欧洲小国卢森堡);⑥事件中主体最后变成的状态(水凝结成了冰)
连事 ^[5-7]	袁鲁刘	除领属关系和类同关系之外的关联关系的相关者,主要包括:①临时性关系的相关者(这栋别墅靠近天目湖);②常识性关系的相关者(广东省毗邻香港);③人际关系的相关者(女儿与继母很和睦);④事件中的牵涉者(世界性经济危机又波及香港)
原因 ^[6]	鲁	造成某种结果或者另一件事情发生的条件(①吴孟达因肝癌于2021年2月去世;②王刚因为贪玩而被老师批评)
依据 ^[6]	鲁	事件发生所遵循的根据或标准(①法庭依照刑法第三条判决王敏死刑;②秘书按照领导的指示办事)
意图 ^[6]	鲁	事件主体或客体所要达到的目的(①他们为了金钱而偷了这辆汽车);②陈飞宇为成名而努力学习)
时间 ^[5-7]	袁鲁刘	事件所发生的时间点或时期:①事件发生的时间点(中央新闻联播于晚上7点播出);②事件发生的时期(霍金于2018年去世)
范围 ^[5-7]	袁鲁刘	事件所关涉的范围或限定的界限:①事件所涉及的范围(所长就当前局面进行了解释);②事件所限定的界限(英语方面他是专家)
物量 ^[5-7]	袁鲁刘	事件所涉及的物体的数量:①事件中物体的数量(中国国旗的星星有5个);②事件中度量衡的数值量(北京到上海大约1318公里)
时量 ^[5-7]	袁鲁刘	事件本身持续的时段或迄今所经历的时段数值。①事件本身所持续经历的时段数值(这篇论文写了一个月还没写完);②事件结束后到现在所经历的时段数值(这袋薯片已经过期了半年)
频量 ^[5-7]	袁鲁刘	事件中行动或变化所进行的次数以及以所用工具表示的动量:①事件中实体行动的次数(我一年去北京四趟);②事件中事物变化的次数(蛇大约每年可以蜕皮4次);③事件中以所用工具表示的动量(实习护士扎了三针才扎到我的血管)

注:本文未对其进行修改,以“判断标准-相应例句”的形式进行总结。

3.2 新提出的和重定义的主要周边语义角色

本节将详细解释新提出和重定义的主要周边语义角色,其被提出和重定义一方面是基于对现有语义角色分类体系^[4-7]和 CSC 语料的仔细分析和思考,另一方面也是基于利用 FSTD 进行 Web 文本语义分析与知识抽取的实际需求^[10-11]。为了能客观地区分和理解这些新提出和重定义的主要周边语义角色,提高标注一致性,本节给出了它们被提出和重定义的原因,并以“判断标准-相应例句”的形式进行说明。

雇施事(Hiring Agent):在对 CSC 进行分析的过程中,发现了一个难以用现有语义角色体系解释的现象,这使得本文提出了一个新的主要周边语义角色。以句子 1 为例。

句子 1:我叔叔建造了一栋别墅

对于句子 1,乍看上去,首先想到事件主体(即

我叔叔)的语义角色是施事。但是仔细思考后,本文意识到“我叔叔”这个句子成分被理解为施事是不恰当的,因为事件中的“我叔叔”自身根本不是建筑者,他不能完成建筑工作,因此“我叔叔”不符合施事的定义。为了解决这个问题,本文引入了一种新的语义角色,称为雇施事。与施事类似,雇施事被定义为有意志的主体,但它不是事件中行为的直接发起者,而是只授予或雇用一些真正的实体来发起事件中的行为操作。

雇施事在事件语义推理中有着重要作用。以句子 1 为例,如果“我叔叔”被标注为“施事”,那就意味着别墅是我叔叔本人亲自建造的。然而,如果“我叔叔”被贴上了“雇施事”的语义角色标签,则根据雇施事的定义,可以进行以下推理:

建造(我叔叔,别墅) & 被标注(我叔叔,雇施事) → ∃x: 建筑队雇用(我叔叔, x) & 建造(x, 别墅)

综上所述,可以推断出在事件中存在一个没有被提及的实体(致事)建筑队,而别墅的实际建造者就是丢失的实体(致事)。

代施事(Resulting Agent):在对 CSC 进行分析的过程中,本文发现了一个不同于施事和雇施事的新的主要周边语义角色,以句子 2 为例说明。

句子 2: 电脑病毒窃取了我们的敏感信息

在句子 2 中,事件主体(即电脑病毒)很容易被标注为施事。然而,电脑病毒是由人类制造的,它们并不是有意志的主体,无法自主地发出相关行为活动,而是在人们的控制下进行窃取敏感信息活动,这不符合施事和雇施事的定义。为了解决该问题,本文引入了一个新的语义角色,称为代施事。代施事被定义为一个有意志的实体的衍生物,它能够代替该有意志的实体本身发出相关行为。为了能够客观地识别代施事,本文总结了三个判断代施事的标准,如表 4 所示。

表 4 代施事的判断标准

判断标准	相应例句
有意志的实体产生的思想、观点或著作	① 这篇文章也确实蛊惑了一批人; ② 法轮功害惨了无数家庭
有意志的实体产生的举措或行为	① 这一措施最终缓解了伦敦严重的交通堵塞; ② 丈夫的冷漠深深伤害了她的心
有意志的实体产生的具有某种特定功能的无意志的实体	熊猫病毒破坏了系统硬盘中的数据文件

根据代施事定义,一旦事件中某个成分被标注为代施事,则可知该角色不是一个有意志的主体,但它能够代替相关有意志的主体执行相关行动。

在鲁川^[6]的工作中,区分受事和客事的关键在于判断客体在事件中是否发生改变。然而,判断客体在事件中是否发生改变是比较主观的,如引言所述,这也造成了 CSC 中语义角色“客事”和“受事”的标注错误率较高。为了能够更好地区分受事和客事,在考察了鲁川^[6]和王亚^[9]的中枢语义角色分类体系后,本文为这两个语义角色分别总结了若干判断标准。

受事(Patient):指的是受事件行动所改变的直接客体。这种客体的共同点在于因主体的行为而受到影响,主体的行为一般包括搬移、交易、管理和控制等。基于这样的思路,并参考现有相关语义角色分类体系^[4-7]以及对 CSC 语料分析后,本文总结了

判断受事的两个标准,如表 5 所示。

表 5 受事的判断标准

判断标准	相应例句
在①形态②位置③所有权④使用权等方面所改变的直接客体	① 爷爷抚摸着这张椅子; ② 张主任从北京带了一些特产; ③ 妈妈送给我一本书; ④ 我在美国租了一辆车
主体执行管理、控制等支配行为 ^[6,9] 所涉及的客体	① 老师处罚了这个淘气的学生; ② 马云掌控着阿里巴巴公司

客事(Entity):指的是事件涉及的但未改变的客体。这种客体的共同点在于不会在因主体的行为而受到影响和变化,主体的行为一般涉及情感、思想、感知和传递信息等。基于这样的想法,并参考现有相关语义角色分类体系^[4-7]以及对 CSC 语料分析后,本文总结了判断客事的六个标准,如表 6 所示。

表 6 客事的判断标准

判断标准	相应例句
感知所获得的信息	① 学生们听到了鸣炮声音; ② 记者看到了感人的一幕
情感所涉及的客体	① 在外求学的我很想念父母; ② 我爱我的祖国
思想所涉及的客体	① 王主任在认真思考问题; ② 我猜出事情的真相;
传播行为 ^[6,9] 所传递的信息	① 鲁老师给学生讲知识工程; ② 公诉人宣读了起诉书
探询所涉及的信息	① 杨天打听仇人的下落; ② 医生测量弟弟的身高
遭受 ^[6,9] 等内向活动所涉及的客体	① 这个女孩收到一束鲜花; ② 我挨了一顿骂

变事(Change):在对 CSC 分析过程中,本文发现了一个受事件行动所改变,但不同于受事的客体。以句子 3 为例进行分析。

句子 3: 王阿姨拧紧螺丝

对于句子 3,事件的客体(即螺丝)很容易被标注为受事。然而,本文随后意识到这个客体不同于受事(“王阿姨拧螺丝”中的“螺丝”),因为句子 3 的中枢语义角色是拧紧,可以推断出王阿姨拧了螺丝,并且螺丝的状态是被拧紧的。而从句子“王阿姨拧螺丝”中,只能知道王阿姨拧了螺丝,至于螺丝的状态是被拧紧、拧松还是其他状态,从句子中无法得到。为了解释这种语义现象,将其与受事进行区分,本文引入了一个新的语义角色,称为变事。经过对

CSC 语料的分析,并考察了鲁川^[6]和王亚^[9]的中枢语义角色分类体系后,发现它们定义的致变类行为可以很好地解释这种语义现象,本文总结了两个判断变事的标准,如表 7 所示。

表 7 变事的判断标准

判断标准	相应例句
致变类行为 ^[6,9] 导致其自身发生改变的客体	① 修理工拧紧了 <u>螺丝</u> ; ② 学生们在操场上撬动了 <u>巨石</u>
致变类行为 ^[6,9] 导致在时间上有所进展的客体	① 我实现了自己的 <u>愿望</u> ; ② 淘宝店家推迟了 <u>买家的订单</u>

虽然受事和变事都能体现受事件行动而改变的客体,但相较于前者,后者蕴含更细致的语义信息:在受事所具有部分语义信息的基础上,变事所在事件的中枢语义角色还能更多地强调导致客体变化后的状态。

源事(Origin):文献[4-7]将时间和空间的起点以及事物发展过程的起始阶段称为源事(源点),这意味着可以从事件中追溯到事件的起源信息。例如,从句子“班长从 3 点到 6 点一直在学习”中可以得出班长学习的起源时间是 3 点。通过对 CSC 进行分析,本文发现除时间、空间和事物发展过程具备起源信息外,主体退出的组织与活动、交际活动^[6,9]中的信息与转移活动^[6,9]中的物品也具备类似的起源信息。例如,从句子“海关工作人员缴获犯罪分子 3.3 公斤海洛因”中可以追溯到“转移活动物品”海洛因来源于犯罪分子,即“海关工作人员从犯罪分子处缴获 3.3 公斤海洛因”。基于这样的原因,本文把主体退出的组织与活动、交际活动和转移活动中作为来源的邻体归纳到源事中。本文总结了 6 个判断源事的标准,如表 8 所示。本文的源事具有“起源”的语义特征,通常可以通过直接或间接的方式由句式“从……”引出^①。

表 8 源事的判断标准

判断标准	相应例句
事件中作为时间的起点(时间的起点)	班长从 <u>3 点</u> 到 6 点一直在学习
事件中作为空间的起点(空间的起点)	张明从 <u>南京</u> 经过天津回到北京
事件中事物发展过程的起始阶段(发展过程的起点)	图像检索方法从特征手工提取发展到 <u>深度学习</u>
事件中主体退出的组织或者活动(组织或者活动的起点)	① 英国正式退出 <u>欧盟</u> ; ② 两名队员因伤退出了 <u>比赛</u>

续表

判断标准	相应例句
事件中交际活动 ^[6,9] 作为来源的邻体(信息的来源)	① 他从 <u>招生办</u> 得知了招生计划; ② 男朋友从 <u>闺蜜</u> 那打听我的事
事件中转移活动 ^[6,9] 作为来源的邻体(物品的来源)	① 海关工作人员缴获 <u>犯罪分子</u> 3.3 公斤海洛因; ② 歹徒抢劫了 <u>银行</u> 所有的现金

宿事(Destination):文献[4-7]将时间和空间的归宿点以及事物发展过程的结束阶段称为宿事(终点)。因此,可以从事件中追溯到事件的归宿信息。例如,从句子“张明从南京经过天津回到北京”中可以得出张明的空间归宿点是北京。通过对 CSC 进行分析,本文发现除时间、空间和事物发展过程具备归宿信息外,主体加入的组织和参加的活动也具备类似的归宿信息。例如,从“王小刚加入了共产党”中可以知道王小刚的组织归宿点是共产党。基于这样的原因,本文将主体加入的组织和参加的活动归纳到宿事,并定义宿事为事件中的主体进行一系列活动后的归属方。为了能容易地辨别宿事,表 9 列出了 5 个判断宿事的标准。本文的宿事具有“归宿”的语义特征,通常可以直接或通过间接转化的方式由句式“至……”引出。

表 9 宿事的判断标准

判断标准	相应例句
主体经过活动后到达的空间归宿点	张明从南京经过天津回到 <u>北京</u>
主体经过活动后所在的时间归宿点	张老师工作到 <u>深夜 11 点</u>
事件中事物发展过程的结束阶段(发展过程归宿点)	图像检索方法从特征手工提取发展到 <u>深度学习</u>
主体加入(Join)的组织(组织归宿点)	王小刚加入了 <u>共产党</u>
主体参加(Participate)的活动(活动归宿点)	马玉梅参加了 <u>蓝桥杯决赛</u>

向事(Target):文献[5-7]将主体移动趋向的邻体称为向事,通过它们可以了解事件的指向目标。例如,从“志愿者向陕南灾区跑去”中可以得到志愿

① 间接方式引出指的是可以将句子转换为“从……”的句式,例如,句子“海关工作人员缴获犯罪分子 3.3 公斤海洛因”可转化为“海关工作人员从犯罪分子处缴获 3.3 公斤海洛因”的“从……”句式。

者移动的指向目标是陕南灾区。通过对 CSC 进行分析,本文发现除了主体趋向的邻体外,交际活动^[6,9]中信息和转移活动^[6,9]中转移物的接收者以及主体行为产生明显利益的接受者也具有类似的目标指向性。例如,“王校长向学生致欢迎词”表明王校长发出欢迎词的指向目标是学生,“美国向日本投放原子弹”表明美国投射原子弹的指向目标是日本。基于这样的原因,本文把事件主体所趋向的邻体、交际活动中信息(客体)或者转移活动中转移物(客体)的接受者以及主体行为产生明显利益的接受者归纳到向事。为了能够容易地辨别向事,表 10 总结了 4 个向事的判断标准及其相关例句。本文的向事具有“指向目标”的语义特征,通常可以直接或通过间接转化的方式由句式“向……”引出。

表 10 向事的判断标准

判断标准	相应例句
事件主体所趋向的邻体(移动主体的指向目标)	① 志愿者向 <u>陕南灾区</u> 跑去 ② 我前往 <u>北京</u>
交际活动 ^[6,9] 中的信息(客体)接受者(发出信息的指向目标)	① 王校长向 <u>学生</u> 致欢迎词; ② 姐姐祝 <u>小刚</u> 生日快乐
转移活动 ^[6,9] 中的转移物(客体)的接受者(转移物的指向目标)	① 美国向 <u>日本</u> 投放原子弹; ② 政府为 <u>志愿者</u> 颁发证书
主体行为产生明显利益的接受者(利益行为的指向目标)	① 年青人给 <u>孕妇</u> 让座; ② 护士为 <u>病人</u> 打针

用事(Reliance):指的是事件中所用到的方法、时耗、工具以及材料等。在现有语义角色分类体系中^[5-7],按照客体种类的不同,用事一般可被分为:①方式类用事;②时耗类用事;③工具类用事;④材料类用事等。然而,在对 CSC 进行分析的过程中,本文发现了一个现有语义角色体系忽略的现象,即用事对应的主体既可以是施事,也可以是当事。以句子 4 和句子 5 为例进行说明。

句子 4:妈妈使用炊具做饭

句子 5: Fonepad7 采用了双前置扬声器设计

句子 4 中用事(即炊具)对应的主体是施事(即妈妈),而句子 5 中用事(即双前置扬声器设计)对应的主体是当事(即 Fonepad7)。因此,按照事件主体类型的不同,本文将用事分为:①施事型用事;②当事型用事。

位事(Location):指的是事件发生所处的空间。文献[4-5,7]将位事理解为事件发生所处的场所,其更多表示物理位置,但在实际语料中存在表示非物

理位置的虚拟空间。文献[6]将位事进一步解释为物理空间和社会空间,但社会空间的概念过于宽泛,不利于理解。为了解决这个问题,本文将位事细化为 5 种常见的空间,即物理空间、网络空间、组织空间、思想空间和其他空间等,位事的判断标准如表 11 所示。

表 11 位事的判断标准

判断标准	相应例句
事件发生所处的物理空间	我在 <u>食堂</u> 吃晚饭
事件发生所处的网络空间	张经理在 <u>微博</u> 发表声明
事件发生所处的组织空间	他在 <u>党组织</u> 学习党的先进知识理论
事件发生所处的思想空间	灰太狼在 <u>懒羊羊的梦境</u> 中抓羊
事件发生所处的其他空间(具备空间性质的其它场所)	杨过在 <u>困境</u> 领悟了新的剑法

在现有语义角色分类体系^[4-7]中,目前没有一个语义角色可以具体地描述事件中对象的某个方面。然而,在对 CSC 进行分析的过程中,发现存在一些描述事件对象的语料。为此,本文提出了三个新的主要周边语义角色。下面以句子 6 和句子 7 为例说明。

句子 6:大海的颜色是蓝色的

句子 7:大海是蓝色的

在上述例句中,蓝色描述大海的某个方面。在基于 FSTD 设计的计算机问答系统中,如果系统被问到大海是什么颜色的?问答系统可以准确地回答句子 6 的问题,但无法回答句子 7 的问题,因为在句子 7 中,没有明确的词来表示颜色。针对该问题,本文引入了三种新的语义角色,即属性、性质和值事。

属性(Attribute):被定义为事件中对象的某一个方面的信息。

性质(Property):被定义为事件中对象本身所具有的与其他对象不同的特征。需要注意的是,性质和属性很容易被混淆,但它们的知识表示方式不同,对于事件中的某个对象,性质通常可以被表示为:性质(对象)。

值事(Attribute-value):指的是事件中一个对象在某个属性上的取值。在知识表示方面,属性和值事通常需要搭配使用,事件中的某个对象通过属性和值事可以表示为:属性(对象,值事)。

根据属性、性质和值事的定义,可以通过以下方式对句子 6 和句子 7 进行语义角色标注:

义文法使用的主要周边语义角色以鲁川语义角色分类体系为主,与本文的主要周边语义角色分类存在差异。此外,为了防止标注人员参考 FSTD 语义文法的主要周边语义角色进行标注,我们把 FSTD 文法中所有主要周边语义角色类型替换成字符串“主要周边语义角色”,使标注人员只能参考中枢语义角色和第 3 节介绍的主要周边语义角色判断标准完成主要周边语义角色类型的判定,保证标注人员在语义角色标注时不受其他因素的影响。

4.2 复合动词作中枢语义角色的处理情况

在对 CSC 进行分析的过程中,本文发现复合动词作中枢语义角色会影响标注人员对主要周边语义角色类型的判断,这类现象目前并没有体现在 FSTD 语义文法中。以句子“红军主力退守宁化县”为例,中枢语义角色是“退守”,“红军主力”是“退守”动作的发起者,因此“红军主力”是施事(施事判断标准①),然而,“退守”是复合动词,它的释义为“后退并防守”,如果理解“退守”的语义中心偏向“后退”,则“宁化县”为宿事(宿事判断标准①),它解释了“红军主力”现在所处的位置;如果理解“退守”的语义中心偏向“防守”,则“宁化县”被标注为受事(受事判断标准⑤),它解释了“宁化县”是被防守的对象。针对此类复合动词作中枢语义角色影响主要周边语义角色类型的句子,为提高语义角色标注的一致性,本文规定理解其中枢语义角色的语义重心偏向于后一个字词。例如,“红军主力退守宁化县”中“退守”的后一个字词为“守”,因此“宁化县”优先被标注为“受事”,同时为这个句子打上“*”标记。复合动词作中枢语义角色影响主要周边语义角色类型的语义现象很有意思,后期我们将对其进行深入研究。

值得注意的是,并非所有的复合动词作中枢语义角色都会影响主要周边语义角色类型的判断。例如,对于句子“项羽攻占了咸阳”,中枢语义角色“攻占”的释义为“攻打并占领”,然而,理解“攻占”的语义中心无论偏向“攻打”还是“占领”,“咸阳”的主要周边语义角色类型都是受事。

4.3 容易混淆的主要周边语义角色难点分析

由于自然语言语义理解的主观性和复杂性,仍然可能会存在部分主要周边语义角色的类型判断让人困惑。本节中对部分容易混淆的典型主要周边语义角色进行难点分析和判断规定。

施事 VS 代施事 VS 当事:事件中的主体发出相关行为时,施事、代施事和当事有时候会混淆。①施事一般强调由有自我意识的主体发出相关可控行为,其主体一般是人、动物或者组织机构等,如“妈妈做饭”中的妈妈。②代施事强调的是由有自我意识的实体所产生的衍生物发出相关行为,这些衍生物没有自我意识,其一般是有意志实体产生的思想、观点、著作或创造的有特定功能且无意志的物品,如“电脑病毒盗取用户隐私”中的“电脑病毒”。③当事更多强调没有自我意识的主体发出相关非可控行为,与代施事不同,这些主体一般是大自然存在的无自我意识的实体,如“黄河向东流去”中的“黄河”。

变事 VS 受事:变事和受事有时容易混淆。变事所在事件的中枢语义角色在鲁川^[6]和王亚^[9]的中枢事元分类体系中被成为“致变类行为”,其被定义为主体导致客体发生化学、物理或生理变化之行动,主要分为:①既表述主体行动又表述客体变化,比如“拧紧”“打倒”“切断”“扩大”“缩小”“打碎”等。②不表述主体行动,只表述客体的变化,比如“丰富”“稳定”“实现”“繁荣”等。而作为对比的受事所在事件的中枢语义角色没有这种特点,其通常只能体现出主体的行动,无法体现客体的变化。

源事 VS 受事:源事在转移活动中作为来源的邻体时,有时会与受事混淆。例如,对于“歹徒抢劫了银行”和“歹徒抢劫了珠宝”这两个句式很相似的句子,很容易把“银行”和“珠宝”都理解为受事。事实上,目前 Chinese Proposition Bank 就把“银行”和“珠宝”都标注为受事(AGR1),然而,这两个句式相似的句子在语义上表达的信息不同,图 5 展示了这两个句子在全息事件网络(Holographic Event Network, HEN)上的映射差异^①。

从图 5 的 HEN 的映射差异可看出“歹徒抢劫了银行”表达的语义是歹徒从银行拿走了很多有价值物品(银行拿不走且不属于歹徒),而“歹徒抢劫了大量珠宝”表达的意思是歹徒拿走了珠宝(珠宝被拿走了且属于歹徒),这导致了在基于 FSTD 设计的

① 注:HEN 是我们课题组在曹存根研究员的主持下正在研发的一种事件网络,其致力于语义分析、知识抽取和常识获取等研究打下基础,HEN 主要包含过程事件网络层和状态事件网络层,其中状态事件网络层类似于一个常识图谱,将实体、概念、属性(值)作为节点,而节点之间的连线被标识为节点之间的关系,这些关系的种类主要基于 ConceptNet5 中的关系而确定。HEN 不是本文工作重心,在此只简要说明,后续课题组会发表相关工作进行详细介绍。

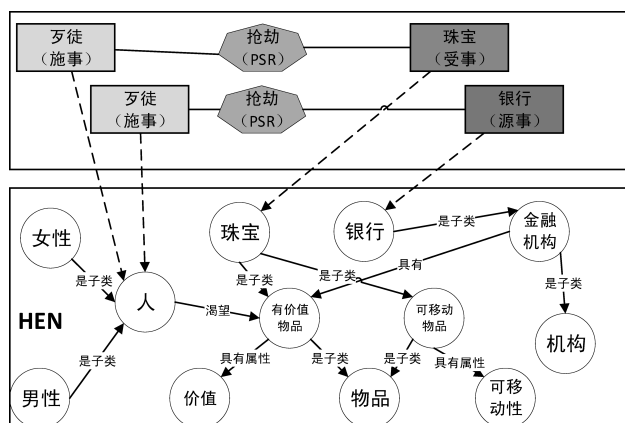


图5 “歹徒抢劫银行/珠宝”在 HEN 上的映射差异

计算机问答系统中,如果输入“歹徒抢劫了银行”,问计算机“银行现在属于谁?”。鉴于“歹徒抢劫了珠宝”中受事“珠宝”现在已属于施事“歹徒”的设计,计算机同样会回答受事“银行”属于“歹徒”,这种回答显然是错误的。对上述两个句子做进一步分析,将“歹徒抢劫了银行”完善为“歹徒抢劫了银行大量珠宝”,则“珠宝”作为被抢走的对象,其所有权和位置发生了改变,因此“珠宝”是受事,此时将“银行”理解为“受事”就无法体现整个句子的语义:珠宝来自银行,歹徒从银行中抢走了珠宝,因此,“银行”理解为源事更合理。为提高标注的一致性,本文规定转移活动中涉及转移物来源时,将其标注为源事。

原因 VS 依据:引起事件的原因和事件所依照的根据是否可以归为一类语义角色在现有文献[6,27]中存在争议。一方面,鲁川^[6]将引起事件的原因和事件所依照的根据归纳到缘由这一语义角色中,其理由在于这两类都能体现出造成某种结果或者另一件事情发生的条件,缺失了这种条件,人们会对产生的结果或者发生的事件提出疑问。例如,对于句子“吴孟达因肝癌于2021年2月去世”,如果缺少“癌症”这个条件,则人们会产生“为什么吴孟达会去世”的疑问;对于句子“法庭依照刑法第三条判决王敏死刑”,如果缺少“刑法第三条”这个条件,则人们会产生“为什么法庭能判决王敏死刑”的疑问。另一方面,邢丹^[27]认为原因和依据在句式上存在较大的差异,其差异主要体现在介词结构搭配的不同,因此,邢丹将引起事件的原因归为语义角色原因,而将事件所依照的根据归为语义角色依据,并总结了两者不同的介词结构搭配:原因的常用搭配介词有由于、由、因、因为等,而依据的常用搭配介词有根据、依据、依、遵照、依照、按、按照、基于等。鲁川和邢丹的观点都有一定的道理,考虑到句式上存在的差异,

本文选用了邢丹^[27]的划分标准,将原因和依据作为两个独立的语义角色理解,但也要意识到原因和依据之间的相关性,鲁川的观点在某些特定的问答系统或者因果关系的常识获取研究中能够帮助获取更多的实验数据。

依据 VS 用事:当依据表示事件发生所遵循的根据和用事表示用到的方式时,两者容易混淆。一般而言,当依据表示事件发生所遵循的根据时,具有明显的参考依照性,是有理可循的,不是随意的;而用事表示方式的时候,仅仅只能体现出使用了某种方式。例如,对于句子“法庭依照刑法第三条判决王敏死刑”中的刑法第三条也可以被理解为法庭判处王敏死刑的方式,但是“刑法第三条”前的“依照”更能体现出法庭判决王敏死刑是有理可循的,并不是随意就使用了刑法第三条就判处王敏死刑的,当问法院为什么能判王敏死刑的时候,法院可以回答其参考依据是“刑法第三条”。而对于句子“县官动用私刑拷问囚犯”这句话中的“私刑”仅仅只能体现县官拷问囚犯的方式,至于这种方式是否有理可循,从“动用”二字是无法体现的。为提高标注一致性,本文规定当句子中的某个成分可以同时被理解为依照的根据和使用的方式时,考察该成分的前一个引导词,如果这个引导词是类似于“依照”“按照”“遵照”“根据”等具有明显参考依照性的词,将这个句子成分标注成依据,而如果这个动词是类似于“使用”“用”“采用”等只具有使用性并不能体现参考依照性的词,将相应的句子成分理解为用事。

经事 VS 用事:当经事表示经历的过程并且用事表示用到的方式时,两者容易混淆。例如,对于句子“小明经过高考考上了大学”,“高考”既可以被理解为小明考上大学所经历的过程,也可以被理解为小明考上大学所用的一种方式。为提高标注一致性,与处理依据和用事的方式类似,当句子中的某个成分可以同时被理解为经历的过程和使用的方式时,本文规定该成分的前一个引导词如果是类似于“经过”“经历”“通过”等具有明显过程经历性的词,将该成分理解为经事。而如果这个类似于“使用”“用”“采用”等只具有使用性并不能体现经历性的词,将相应的句子成分标注为用事。

4.4 辅助周边语义角色的标注规定

在对 CSC 进行语义角色标注的实践中,发现标注人员在标注辅助周边语义角色“受益者”和“受损者”时存在立场问题,例如,对于句子“红军攻占阳新

县城”，其句子成分“红军”和“阳新县城”分别是主要周边语义角色施事和受事。然而，在为“阳新县城”标注辅助周边语义角色时，标注人员有了分歧：部分标注人员标注“阳新县城”是受损者，他们认为“阳新县城”是被攻击方，被攻击代表自身价值受损。另一部分标注人员标注“阳新县城”是受益者，他们给出的理由是“红军”代表正义的一方，占领阳新县城能够解放阳新县城的百姓，使他们不再被统治者所统治。事实上，标注人员产生分歧的根本原因在于他们站在不同的立场考虑问题，认为“阳新县城”是受损者的标注人员所处的立场是阳新县城现在统治者的立场，阳新县城被占领造成了统治者的损失。而认为“阳新县城”是受益者的标注人员所处的是阳新县城百姓和红军的立场，阳新县城被占领能够解放阳新县城被剥削的百姓，红军能够给这些百姓更好的生活。为提高标注的一致性，本文规定，在进行辅助周边语义角色标注时，对于因为标注立场不同而存在分歧的句子，采用绩效风险的方式进行处理，即标注人员要站在中立者的角度进行辅助周边语义角色的标注，在没有任何确定信息的情况下，为

了不犯错，对无法确定受益者还是受损者的句子成分不进行辅助周边语义角色的标注。

以上为提高语义角色标注一致性进行了难点分析说明和判断规定，以便帮助标注人员进行理解、抉择和判断。随着后续大规模语义角色标注工作的展开，我们可能还会遇到目前语义角色分类体系没有考虑的问题，届时我们将对其进行深入分析，积极与其他专家交流，将所遇到的问题进行总结，不断地完善我们的语义角色分类体系。

5 标注实践与分析

为了更好地进行语义角色的标注工作，我们基于 Doccano^① 和鲁棒 Earley 语义解析器^[10] 开发并部署了一个语义角色标注平台。利用这个语义角色标注平台，我们邀请了一批标注人员对 CSC 中 9 541 条句子进行了语义角色的重新标注。标注人员由具有计算机和语言学专业背景的在读研究生组成，并且在标注前经过了两周左右的培训。语义角色标注流程如图 6 所示。

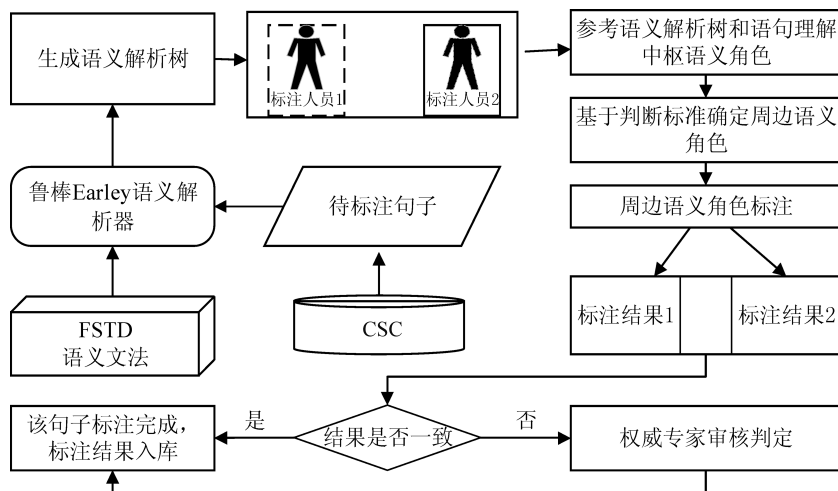


图 6 语义角色标注流程

如图 6 所示，为了保证语料中语义角色的标注质量，CSC 中的每一条句子均由两位标注人员独立标注。如果两位标注人员标注句子成分对应的语义角色标签都一致，则将该句子确定为标注完成，否则请权威专家进行分析和判定。为了便于管理和统计，我们将 CSC 中的语料划分成 5 组，每组约 1 908 条句子，CSC 语义角色的标注工作以组为单位，目前 5 组数据的标注工作已经全部完成。本文用 Kappa 系数^[28] 定量地衡量语义角色标签标注的一致

性，具体每组数据标注的 Kappa 系数如图 7 所示。

从图 7 中可以看出，对于每组数据，两位标注人员标注主要周边语义角色标签和辅助周边语义角色标签的 Kappa 系数普遍维持在 0.78 以上，5 组数据主要周边语义角色标签标注的平均 Kappa 系数为 0.800 4，辅助周边语义角色标签标注的平均 Kappa 系数为 0.862 0，表明人工标注具有较好的一致性，达到了预期目标。

① <https://github.com/doccano/doccano>

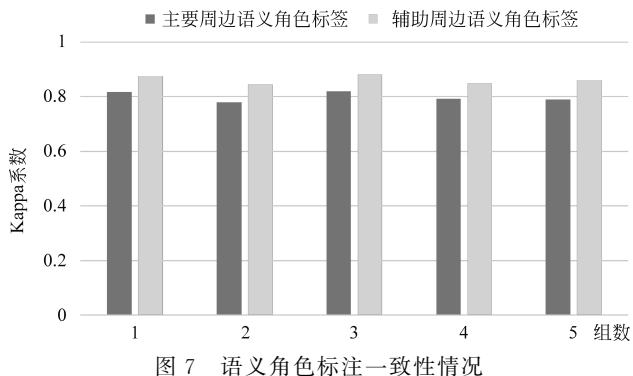


图7 语义角色标注一致性情况

表12统计了本文新提出和重定义的主要周边语义角色在CSC中的分布情况。参与语义角色分布情况统计的主要周边语义角色包括代施事、受事、客事、变事、源事、宿事、向事、用事、位事、属性、性质和值事。由于雇施事语义性较强，在语料中能够明确被标注为雇施事的相关语句数量较少，因此雇施事不进入统计。

表12 主要周边语义角色在CSC中的分布情况

主要周边语义角色	被标注的语料数量	语义角色出现率/%
代施事	161	1.69
受事	1 813	19.00
客事	1 354	14.19
变事	541	5.67
源事	184	1.93
宿事	232	2.43
向事	644	6.75
用事	63	0.66
位事	203	2.13
属性	44	0.46
值事	44	0.46
性质	42	0.44
重定义的语义角色	3 917	41.05
重定义的语义角色(除受事和客事外)	1 128	11.82
新提出的语义角色	781	8.19

由表12可以看出，本文新提出和重新定义的主要周边语义角色在语料中占据较大的比例，其中重定义语义角色的出现率为41.05%，新提出周边语义角色的出现率为8.19%。前者出现率较高的原因是受事和客事在被标注语料中的出现率比较高，将受

事和客事排除之后，重定义语义角色的出现率为11.82%。表12的统计结果表明了本文新提出和重定义的主要周边语义角色在CSC的语义角色标注过程中发挥着作用。

为了进一步检验数据标注的合理性，我们使用鲁川的语义角色体系^[6]对CSC进行了主要周边语义角色的标注，并与本文提出的语义角色分类体系标注的结果进行了对比，结果显示，对于相同的句子成分，本文与鲁川标注相同语义角色标签的一致率为72.55%。在不一致的27.45%的被标注的语义角色标签中，占全部不同被标注语义角色标签比重较大的标签如表13所示，其他的比重都在1%以下。

表13 相同句子成分、标注不同的语义角色标签的部分比重

本文语义角色	鲁川语义角色	占全部不同语义角色的比例/%
代施事	施事	3.74
代施事	当事	2.41
领事	涉事	5.46
变事	致事	26.37
宿事	受事	1.15
宿事	客事	2.14
向事	受事	3.28
向事	客事	2.86
向事	宿事	1.26
向事	涉事	8.82
连事	涉事	13.02
依据	缘由	1.23
原因	缘由	1.41

在本文与鲁川体系标注结果不同的主要周边语义角色中：

(1) 本文标注为代施事而鲁川语义角色体系标注为施事或当事的比重为6.15%。这是由于本文把有意识的实体产生的思想、观点、著作、举措或具有某种特定功能的无意识的物体理解成代施事，而鲁川体系没有对其进行细致的划分，直接将其理解为施事或当事。代施事的提出有助于解决施事在语料中语义分布不一致的问题^[29]。

(2) 本文标注为领事，而鲁川语义角色体系标注为涉事的比重为5.46%。这是由于鲁川体系对涉事的定义比较模糊且如第2节所述存在矛盾，而本文将涉事剔除了。例如，对于句子“水果店坏了四筐

梨”，鲁川体系将“水果店”标注为涉事，而本文认为该句子应当被理解为“水果店拥有的梨子坏了”，因此水果店是领事，同时，给水果店标注辅助周边语义角色表示水果店是受损者。此外，在鲁川体系中被标注为涉事，而在本文体系中被标注为其他语义角色的有向事和连事，其占比分别为 8.82% 和 13.02%。这是因为本文将交际活动中信息的接受者、转移活动中转移物的接受者理解为向事，并将事件中的牵涉者理解为连事。

(3) 本文标注为变事而鲁川语义角色体系标注为致事的比重为 26.37%。这是因为鲁川体系把事件中促使类行为和致变类行为相关的客体都理解为致事，而本文认为这种方式过于粗糙，不利于 Web 文本语义分析和知识抽取^[10,11]，因此将致变类行为导致客体发生变化并能够体现客体状态的句子成分理解为变事。

(4) 本文标注为宿事而鲁川语义角色体系标注为受事或者客事的比重为 3.29%。这是因为本文把主体加入的组织和主体参加的活动均理解为主体进行一系列活动后的归属方，即宿事，而单纯地将其理解为受事或者客事并不能体现这种语义信息。

(5) 本文标注为向事而鲁川语义角色体系标注为受事、客事或者宿事的比重为 7.4%。这是因为本文把事件中信息客体和转移物客体的接受者标注为向事，这样的理解方式有助于将宿事与向事进行区分。

(6) 本文标注为依据或原因而鲁川语义角色体系标注为缘由的比例为 2.64%。这是因为本文将鲁川体系缘由中的依据和原因进行了区分。在本文体系中，引起事件某种结果的条件或引起另一件事发生的条件才被称为原因，而事件发生所依照的根据或依托被标注为依据。

通过对比分析本文和鲁川体系在 CSC 语料上的标注结果，可以发现两者的相同与不同之处。因为 FSTD 语义文法是以鲁川语义角色分类体系为基础而设计的，这样的对比分析有助于快速、高效地建立基于本文体系的 FSTD 语义文法。由于语义理解的复杂性，本文的语义角色分类体系可能存在未发现的问题，再加上部分标注人员对本文语义角色分类体系的理解程度不够深入，所以对 CSC 语料语义角色的标注结果可能也存在一定的错误，但总体来看，语义角色分类和标注的效果达到了预期目标。未来，我们将基于本文的语义角色分类体系进行更大规模语料的语义角色标注。

6 结束语

本文通过引入新的语义角色、吸收和重定义现有语义角色体系中的语义角色的方式，提出了一种改进的汉语语义角色分类体系。为了减少语义理解的主观性以便客观地进行语义角色标注，本文从语义和句式的角度以“判断标准-相应例句”的形式详细地解释了语义角色分类体系中的主要周边语义角色，并进行了提高语义角色标注一致性的说明。最后，通过对语料库 CSC 进行了语义角色的标注实践，本文分析了数据上的标注一致性，统计了新提出和重定义的主要周边语义角色在语料库中的分布情况，并与使用鲁川语义角色分类体系标注的结果进行了对比分析。

尽管本文为提高语义角色标注一致性做了很多努力，例如第 3 节中主要周边语义角色的判断标注和相应例句，第 4 节为提高语义角色标注一致性而做的规定，但由于语义具有一定的模糊性，人们在理解中有一定主观性，受个人知识结构局限，会导致有一定的不一致性，不同的学者对本文的语义角色分类体系仍然可能会有不同的看法，这是目前语义角色分类体系没有得到统一的重要原因，也是自然语言语义理解的难点所在。随着后续对更大规模语料进行语义角色标注工作的进行，我们还会根据标注中遇到的问题和实际任务需求对本文的语义角色分类体系进行不断更新和完善。此外，目前 CSC 中语料数量还不多，后续我们将在标注更多的语料后将 CSC 进行发布公开，以此希望能够获得更多学术界同行的意见和反馈。

参考文献

- [1] WAITE W. Semantic analysis[M]. Berlin: Springer Press, 1974.
- [2] 刘亚慧, 杨浩苹, 李正华, 等. 一种轻量级的汉语语义角色标注规范[J]. 中文信息学报, 2020, 34(4): 10-20.
- [3] MÀRQUEZ L, CARRERAS X, LITKOWSKI K C, et al. Semantic role labeling: an introduction to the special issue[J]. Computational Linguistics, 2008, 34(2): 145-159.
- [4] 朱晓亚. 现代汉语句模研究[M]. 北京: 北京大学出版社, 2010.
- [5] 袁毓林. 基于认知的汉语计算语言学研究[M]. 北京: 北京大学出版社, 2008.

- [6] 鲁川. 知识工程语言学[M]. 北京: 清华大学出版社, 2010.
- [7] 刘茂福, 胡慧君. 基于认知与计算的事件语义学研究[M]. 北京: 科学出版社, 2013.
- [8] ZANG L, WANG W, WANG Y, et al. A Chinese framework of semantic taxonomy and description: preliminary experimental evaluation using web information extraction[C]// Proceedings of the 8th International Conference on Knowledge Science, Engineering and Management, 2015: 275-286.
- [9] 王亚. 基于语义分类的常识知识获取方法研究[D]. 南宁: 广西师范大学硕士学位论文, 2015.
- [10] 方芳. Web 文本语义分析与知识获取方法研究[D]. 北京: 中国科学院大学博士学位论文, 2019.
- [11] 曹聪. 基于语义分类和描述框架的常识知识获取方法研究[D]. 北京: 中国科学院大学博士学位论文, 2015.
- [12] 王亚, 陈龙, 曹聪, 等. 事件常识的获取方法研究[J]. 计算机科学, 2015, 42(10): 217-222.
- [13] QUILLIAN M. "Semantic memory" in semantic information processing[M]. Cambridge: MIT Press, 1968.
- [14] DAVID W. Case grammar: development of the matrix model[J]. Rocky Mountain Review of Language & Literature, 1980, 34(2): 272-273.
- [15] MILLER G, RICHARD B, CHRISTIANE F, et al. Introduction to WordNet: An on-line lexical database[J]. International Journal of Lexicography, 1991, 3(4): 235-244.
- [16] FELLBAUM C, MILLER G. WordNet: an electronic lexical database[M]. Boston: MIT Press, 1998.
- [17] BAKER C, FILLMORE C, LOWE J. The berkeley FrameNet project[C]// Proceedings of the 36th annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, 1998: 86-90.
- [18] CHARLES J, CHRISTOPHER R, MIRIAM L. Background to framenet[J]. International Journal of Lexicography, 2003, 16(3): 235-250.
- [19] SCHULER K, KORHONEN A, BROWN S. VerbNet overview, extensions, mappings and applications[G]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, 2009: 13-19.
- [20] SCHULER K. VerbNet: a broad-coverage, commonality verb lexicon[J]. Dissertation Abstracts International: Sciences and Engineering, 2005, 66(6): 33-45.
- [21] 鲁川. 汉语的意合网络[J]. 语言文字应用, 1998, 26(2): 82-88.
- [22] 王丽杰. 汉语语义依存分析研究[D]. 哈尔滨: 哈尔滨工业大学硕士学位论文, 2010.
- [23] 柏晓鹏. 中文命题库的全局性语义角色标注及其对汉语研究的影响[J]. 语言科学, 2017, 16(5): 481-92.
- [24] 王诚文, 钱青青, 荀恩东, 等. 三元搭配视角下的汉语动词语义角色知识库构建[J]. 中文信息学报, 2020, 34(9): 19-27.
- [25] 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): 15-27.
- [26] 董振东, 董强, 郝长伶. 知网的理论发现[J]. 中文信息学报, 2007, 21(4): 3-9.
- [27] 邢丹, 饶高琦, 荀恩东, 等. 基于大规模语料库的介词结构搭配库构建[J]. 中文信息学报, 2020, 34(11): 1-8.
- [28] CARLETTA J. Assessing agreement on classification tasks: the kappa statistic[J]. Computational Linguistics, 1996, 22(2): 249-254.
- [29] 陈龙, 詹卫东. 施事的语义分布考察与动词的语义特征[J]. 中文信息学报, 2019, 33(1): 1-9.



宋衡(1990—), 博士研究生, 主要研究领域为知识抽取、语义理解。
E-mail: song_heng@yeah.net



王亚(1988—), 博士研究生, 主要研究领域为语义理解、常识知识获取。
E-mail: wangya@ict.ac.cn



曹存根(1964—), 博士, 研究员, 主要研究领域为大规模知识获取、文本挖掘。
E-mail: cgcao@ict.ac.cn