文章编号：1003-0077（2017）00-0000-00

一种细粒度的汉语语义角色标注数据集的构建方法

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：语义角色对自然语言的语义理解和分析有着重要的作用，其自动标注技术依赖良好的语义角色标注训练数据集。目前已有的大部分语义角色数据集在语义角色的标注上都不够精确甚至粗糙，不利于语义解析和知识抽取等任务。为了满足细粒度的语义分析，该文通过对实际语料的考察，提出了一种改进的汉语语义角色分类体系。在此基础上，以只有一个中枢语义角色的语料作为研究对象，提出了一种基于半自动方法的细粒度的汉语语义角色数据集构建方法，并构建了一个实用的语义角色数据集。截至目前，该工程一共完成了 9550 条汉语语句的语义角色标注，其中含有 9423 个中枢语义角色，29142 个主要周边语义角色，3745 个辅助周边语义角色，172 条语句被进行了双重语义角色标注，以及 104 条语句被进行了不确定语义事件的语义角色标注。我们采用 Bi-LSTM+CRF 的基线模型在构建好的汉语语义角色数据集和公开的 Chinese Proposition Bank 数据集进行了关于主要周边语义角色的基准实验。实验表明，这两个语义角色数据集在主要周边语义角色自动识别方面的差异，并且为提高主要周边语义角色的识别准确率提供了依据。

关键词：语义角色；细粒度语义标注；汉语语义角色标注；汉语语义分析

中图分类号：TP391 **文献标识码：**A

Construction of a Finely-Grained Training Dataset for Chinese Semantic-Role Labeling

SONG Heng[1,2], CAO Cungen[1], WANG Ya[1,2] and WANG Shi[1]

(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, and automatically and correctly labeling semantic roles of sentences need a good semantic-role training dataset, but most of the existing semantic-role training datasets are relatively rough or even misleading in labeling semantic roles, which is not conducive to semantic parsing and knowledge extraction. In order to meet the requirements of fine-grained semantic analysis, an improved taxonomy of Chinese semantic roles is proposed by investigating a real-world corpus. On this basis, taking the corpus with only one pivotal semantic role as the research object, we propose a fine-grained Chinese semantic role dataset construction based on a semi-automatic method, and construct a semantic role labeling dataset. Up to now, a corpus containing 9550 sentences has been labeled with 9423 pivot semantic roles, 29142 principal peripheral semantic roles and 3745 auxiliary peripheral semantic roles, which include 172 sentences which are double-labeled with semantic roles and 104 sentences which are labeled with semantic roles of uncertain semantic events. We use a Bi-LSTM+CRF model to conduct experiments on the constructed Chinese semantic-role dataset and the Chinese Proposition Bank dataset. The experimental results show the differences between the two semantic role

收稿收稿日期：2017-03-16；**定稿日期**：2017-04-26

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）。

datasets in automatically recognizing the principal peripheral semantic roles, and provide ideas for how to improve the accuracy of recognizing the principal peripheral semantic roles.

Key words: Semantic role; fine-grained semantic labeling; Chinese semantic role labeling; Chinese semantic analysis

0 引言

语义理解和分析是自然语言处理的重要目标之一，其致力于获取给定文本所蕴含的语义信息，并以计算机能理解的某种方式进行展示[1]。目前，语义理解和分析的研究主要包括深层语义分析和浅层语义分析，其中深层语义分析的相关工作主要包括语义依存分析(Semantic Dependency Parsing, SDP)[2]和普适概念认知标注(Universal Conceptual Cognitive Annotation, UCCA)[3]等。然而，深层语义分析存在语义层次涉及范围广，难以用良好的形式化方法展示所有的语义信息等问题，甚至受限于当前的技术水平，短期内难以形成具有较强实用性的成果[4]。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）[5]，具有多语言通用、表现形式自然、语义结构稳定以及相关模型和算法研究深入等优点。目前，语义角色标注技术被广泛地应用于知识抽取、机器翻译、自动文本摘要、信息检索和自动问答等多种下游任务。

作为实现浅层语义分析的基础，标有语义角色的语料资源至关重要，其可以极大地促进语义分析相关模型和算法的测试与研究。目前，国外比较知名的语义角色数据集资源包括 FrameNet[6]、Proposition Bank[7]和 NomBank[8]等。国内的语义角色数据集有 Chinese Proposition Bank[9]、山西大学汉语框架语义知识库[10]、北京大学中文网库[11]、苏州大学汉语开放谓词论元数据集[4]以及中科院计算所的基于语义分类和描述框架(Framework of Semantic Taxonomy and Description, FSTD)的知识库[13, 22]。相较于国外的语义角色标注数据集，国内的语义角色语料资源增加较为缓慢，且大部分语义角色语料资源没有被公开。受限于较长的标注语料，现有的语义角色标注数据集存在语料标注精度和粒度的问题，主要体现在三个方面：

第一，语料标注所用的语义角色种类不够丰富，导致标注的语料在语义上存在偏差。例如，对于句子“我家安装了百兆宽带”，现有语义角色分类体系将句子成分“我家”标注成施事，这在语义上是不正确的，因为“我家”不是真正安装宽带的主体，而是我家“雇用”了宽带安装人员来进行宽带安装。我们将在第 2 节回顾这个问题。

第二，现有的大部分语义角色数据集标注所用的语料存在多个谓词，但在实际语义角色标注时，只标注语句中的其中一个谓词及其相关语义角色，而其它谓词以及相关的语义角色则未予标注。显然，这种标注方式会丢失语料中那部分没有进行语义角色标注的句子成分的语义信息。我们将在第 3 节回顾这个问题。

第三，现有的大部分语义角色数据集标注方式单一，仅为语料中句子的每个成分标注一个语义角色，而忽略了某些句子成分可能同时扮演多个语义角色。以下面两个句子为例：

例句 1：女警误杀了队友

例句 2：女警枪杀了歹徒

从常规语义角色标注的角度来看，两个句子中的成分“女警”都是施事；但从语义角色受损还是受益的角度来看，例句 1 中的 “女警”是个受损者，而例句 2 中的“女警”是个受益者。又如：

例句 3：八路军主力退守沂蒙山

常规来看，“沂蒙山”是被防守的对象，因此“沂蒙山”是受事，但是从另一层语义去理解，“沂蒙山”是八路军主力到达的地点，则“沂蒙山”还应该是宿事。我们将在第 3 节回顾这个问题。

为了解决上述语义角色标注数据集存在的问题，并更好地满足语义理解分析和知识获取等研究需要[12-15]，本文提出了一种细粒度的汉语语义角色数据集构建方法，并利用该方法初步构建了一个汉语语义角色数据集。本文的主要贡献如下：

（1）基于对已有的语义角色分类体系的分析和实际语料的考察，本文提出了一种改进的汉语语义角色分类体系。该体系将语义角色分为中枢语义角色和周边语义角色，并且将周边语义角色分为主要周边语义角色和辅助周边语义角色。此外，基于细粒度语义分析的需要以及语义角色标注难度的权衡，我们将主要周边语义角色的种类分为 32 种，其中包括 7 个全新的主要周边语义角色以及 5 个常用但经过重新定义的主要周边语义角色。改进的汉语语义角色分类体系解决了现有语义角色数据集语义角色种类不够丰富的问题。

（2）本文提出的细粒度的汉语语义角色数据集构建方法包括中枢和主要周边语义角色标注、辅助周边语义角色标注、语义角色的双重标注以及不确定语义事件的语义角色标注等四个步骤，

其中辅助周边语义角色标注、语义角色的双重标注以及不确定语义事件的语义角色标注解决了现有的大部分语义角色数据集标注方式单一的问题，为我们标注的语料带来更细粒度的语义信息。此外，我们还提出了主要周边语义角色关系约束的概念，有助于设计算法对语料库中标注的语句自动地进行初步审查，减轻后期人工复审的压力。

（3）我们初步构建了一个拥有 9550 条语句的汉语语义角色数据集。相较于现有的汉语语义角色数据集，我们的语义角色数据集中标注的语料拥有更细粒度的语义信息。细粒度的语义信息不仅体现在我们标注的语义角色数据集中拥有更丰富的主要周边语义角色类型，还体现在我们语义角色数据集标注步骤的多样性。

（4）我们采用 Bi-LSTM+CRF 的基线模型在构建好的汉语语义角色数据集和 Chinese Proposition Bank 数据集进行了关于主要周边语义角色的基准实验。我们还分析了基线模型在本文语义角色数据集识别错误的语句，并针对这些识别出错的语句提出了解决这些错误的思路。

本文的组织结构如下。第 1 节介绍了相关研究。第 2 节概述了一种改进的汉语语义角色分类体系。第 3 节详细介绍了半自动的细粒度的汉语语义角色数据集构建方法。第 4 节进行了关于主要周边语义角色的基准实验。第 5 节总结全文并提出未来工作。

1 相关研究

1.1 语义角色分类体系相关研究

为了从语义的角度弥补转换生成语法的不足，菲尔墨（C.J.Fillmore）于 1968 年发表了著作《格辨》[16]，并提出了格语法，他认为“格”能够真正构成自然语言深层结构中的语法关系，最初格的种类主要包括 6 种，分别为施事格、工具格、客体格、处所格、承受格和使动格。后来菲尔墨对其进行了完善，提出了一个含有 13 种格的体系[17]，新增加的格有感受格、源点格、终点格和受益格等。“格”本质上就是本文中的语义角色。

国内学者针对汉语的特点对语义角色及其分类体系进行了大量研究。朱晓亚[18]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，她定义了 14 种动元的语义角色。袁毓林[19]将语义角色称为论元，他总结了现代汉语的 17 种论元，并对这些论元的语义定义和句法特征进行了详细的解释和说明。鲁川[20]将语义角色称为事元，他将事元分为中枢事元和周边事元，并以中枢事元作为现代汉语基本句模的分类标准。鲁川总结了 26 种中枢事元和 26 种周边事元。刘茂福等[21]在认知科学和题元理论的基础上，总结了 16 种原子事件的语义角色类型。由此看出，国内各个学者提出的语义角色分类体系的差异主要在于语义角色类型的数量和具体语义角色术语的定义方面，其本质是语义角色分类粗粒度和细粒度的抉择。

1.2 语义角色相关知识库构建现状

国外比较知名的语义角色相关知识库有 FrameNet、Proposition Bank 和 NomBank 等。FrameNet[6]是美国加州大学伯克利分校构建的基于真实语料库的计算机词典，它以框架语义作为理论基础，能够描述具有相同或相似语义角色的单词。Proposition Bank[7]是美国滨夕法尼亚大学建立的一个集语义词典和标注语料库于一身的论元角色语义知识库，它以动词词典为标注基础，以 Penn TreeBank Ⅱ为标注底层，以动词的论元角色为标注对象。NomBank[8]由美国纽约大学构建的语义知识库，它采用了和 Proposition Bank 大致相同的框架集，几乎涵盖了宾州树库中所有标注过的名词，并详细描述和定义了名词的论元结构。

国内学者对汉语语义角色知识库也进行了大量的研究，主要有 Chinese Proposition Bank、山西大学汉语框架语义知识库、北大网库以及苏州大学汉语开放谓词论元数据集等。Chinese Proposition Bank[9]基本继承了 Proposition Bank 的标注体系，将语义角色分为核心语义角色和附属语义角色。其中，核心语义角色以 Arg0~Arg5 进行标注，附属语义角色以 ArgM 作为前缀进行标注。山西大学的汉语框架语义知识库[10]是以 C.J.Fillmore 的框架语义学为基础，以 FrameNet 为参照的汉语词汇语义知识库，它由框架库、句子库和词元库构成。北京大学中文网库[11]是北京大学袁毓林主持建立的汉语语义关系标注语料库，它在北京大学汉语句法分析树库的基础进行语义标注，北京大学中文网库一共定义了 21 种论元角色。苏州大学汉语开放谓词论元数据集[4]为了达到轻量级的目的，根据句子上下文信息直接标注谓词相关的论元角色，并采用基于词的论元单位表示，避免了论元角色标注任务对谓词语义框架的依赖和对片段边界确定困难的问题。

遗憾的是，目前国内大部分语义角色的相关数据集都没有被公开。本文提出了一种细粒度的汉语语义角色数据集构建方法，相较于现有的语义角色数据集，我们的语义角色数据集中标注的语料具有更细粒度的语义信息，能够更好地满足语义理解分析和知识获取等研究工作的需要[12-15]。

2 一种改进的汉语语义角色分类体系

在本文的汉语语义角色分类体系中，我们首先将语句中的语义角色分为两大类：中枢语义角色（Pivotal Semantic Role, PSR）和周边语义角色（Peripheral Semantic Roles, PSRs），再将周边语义角色分为主要周边语义角色（Principal Peripheral Semantic Roles, PPSR）和辅助周边语义角色（Auxiliary Peripheral Semantic Roles, APSR），下面分别进行介绍。

2.1 中枢语义角色

中枢语义角色表示的是语句的中枢，它在汉语句子中扮演着谓词的角色。中枢语义角色是一个句子的主干，并且一个句子中只能有一个中枢语义角色。本文的中枢语义角色的选取是根据课题组构建的一种基于语义分类和描述框架 FSTD 的知识库[13,22]中的语义类而确定的，其范围主要包括动词和形容词。FSTD 中的语义类是以鲁川[20]定义的 26 种中枢事元为基础，并借助《同义词词林》和《形容词词典》等汉语词典进行补充而完成的。

2.2 周边语义角色

2.2.1 主要周边语义角色

主要周边语义角色表示语句中枢的关键周边语义角色。一个语句中可能存在多个主要周边语义角色。主要周边语义角色类型的确定是语义角色数据集构建过程中最重要也是最复杂的工作之一。我们需要在主要周边语义角色的种类数量方面寻找一个平衡，即在能够保证获得必要的语义信息的同时，尽可能地精简主要周边语义角色的种类。如果主要周边语义角色的种类过少，会造成无法准确地获取文本的语义信息，造成后期知识抽取和问答系统等下游任务无法顺利进行。例如，在“歹徒抢劫了银行”和“歹徒抢劫了珠宝”这两个句式很相似句子中，Chinese Proposition Bank 把“银行”和“歹徒”都理解为受事（AGR1），这是不合理的，因为这两个句子在语义上表达的意思不同：“歹徒抢劫了银行”蕴含的语义是歹徒从银行中拿走了很多银行的贵重物品，而“歹徒抢劫了珠宝”蕴含的意思是歹徒拿走了珠宝，这就导致了在计算机问答系统中，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”现在已经属于施事“歹徒”的设计，计算机同样会回答受事“银行”现在属于“歹徒”，这种回答显然是错误的。而如果主要周边语义角色的种类过多，又会增加标注人员对主要周边语义角色标注的难度。在参考了朱晓亚[18]、袁毓林[19]、鲁川[20] 、刘茂福[21] 等学者以及我们先前设计[13, 22]的语义角色分类体系，并经过实际的汉语语料考察与验证后，我们最终确定了 32 个主要周边语义角色类型（如表 1 所示），包括本文提出的 7 个全新的主要周边语义角色，即雇施事、代施事、变事、空间、属性、性质和值事（在表 1 中以+标记），以及本文重新定义的 5 个常见主要周边语义角色（在表 1 中以*标记），即受事、客事、源事、宿事以及向事。为了在实际语料标注过程中能够准确并容易地区分确定这些主要周边语义角色类型，我们还在表 1 中给出了它们的判断标准以及例句。

表 1 主要周边语义角色种类汇总表

主要周边语义角色		判断标准
主体	施事	发出直接可控行为或思维活动的有意志的主体：①自身亲自执行可控行动的有意志的主体(学生们飞快地奔向食堂)；②发出指令指使有意志的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志的主体(王主任认真思考问题)；
	雇施事+	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：我叔叔建造了一栋别墅（注：此处我叔叔并非别墅的真正建造者，而是其雇佣建筑队建筑别墅的说法。同样的例句还有“我家安装了百兆宽带”等，“我家”也是雇施事）
	代施事+	有意志的实体产生的衍生物，它能够代替该有意志的实体本身发出相关行为，该衍生物主要包括：①有意志的主体产生的思想，观点或著作（法轮功害惨了无数家庭）；②有意志的主体产生的举措或行为（这一措施最终缓解了伦敦严重的交通堵塞）；③有意志的主体产生的具有某种特定功能的无意志的物体等（电脑病毒窃取了我们的敏感信息）；
	当事	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(雨水冲刷着泥土)；②表现自身状态的主体(计算所大楼好高)；③除领属关系[20,22]之外，各种关系的主体(计算所靠近融科资讯中心)；
	感事	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	领属关系[20,22]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)；
客体	受事*	事件涉及的改变的客体，主要包括支配行动在：①形态（爷爷拆了这张椅子）、②位置（张主任从北京带回了一些特产）、③所有权（妈妈送给我一本书）、④使用权（我在美国租一辆车）等方面所改变的直接客体；⑤主体执行管理、控制等行为所涉及的客体（老师处罚这个学生）；
	客事*	事件涉及的未改变的客体。主要包括：①感知所获得的信息（学生们听到了鸣炮声音）；②情感所涉及的客体（我爱伟大的祖国）；③思想所涉及的客体（我猜出了事情的真相）；④传播所传递的信息（公诉人宣读了起诉书）；⑤探询所得到的信息（警方一直在调查通缉犯的下落）；⑥遭受等内向活动所关涉到的客体（我挨了老师一顿骂）；
	成事	主体创建或产生的新客体以及主体实施参与事件产生的无形（Intangible）结果：①有意志主体创建的新客体(联想公司制造了一款新电脑)；②有意志主体对客体进行加工造成客体形成的新状态(面点师傅把面团揉成了面条)；③有意志主体实施或参与事件产生的无形（Intangible）结果(弟弟期末考了 90 分)；④无意志主体产生的新客体(蚕的脑神经分泌一种脑激素)；
	变事+	致变类行为[20,22]造成其自身发生变化或在时间上有所进展的客体：①致变类行为导致其自身发生改变的客体（妈妈不小心打碎了杯子）；②致变类行为导致在时间上有所进展的客体（他们成功地拖延了德军的进攻速度）；
	致事	主体通过促使类行为[20,22]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 (老师安排学生做作业；朝鲜请求中国派兵援助；作者邀请他为这本新书拟写书评)；
邻体	源事*	事件中时空和过程的起点以及作为来源的邻体：①事件中作为时间的起点（班长从 3 点到 6 点一直在学习）；②事件中作为空间的起点（我从南京经过天津回到北京）；③事件中作为过程的起点（图像检索方法从特征手工提取发展到深度学习）；④事件中交际活动[20,22]作为来源的邻体（学生向老师请教一道数学难题；他采访了许多优秀篮球运动员）；⑤事件中转移活动[20,22]中作为来源的邻体（海关工作人员查获犯罪份子 3.3 公斤的海洛因；歹徒抢劫了银行）；

	经事	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)；
	宿事*	事件中的主体进行一系列活动后的归属方：①主体经过活动后已经到达的空间归宿点（我从南京经过天津回到北京）；②主体经过活动后已经达到的时间归宿点（张老师工作到深夜）；③主体加入（Join）的一个组织（项羽投降刘邦；我加入了共产党）；④主体参加（Participate）到的过程（马玉梅献身祖国教育事业；我参加一个会议）；
	向事*	事件所趋向的邻体或信息、转移物以及的有意志或无意志主体行为所产生利益的接受者：①动态所趋向的邻体（志愿者奔赴陕南灾区）；②交际活动中的信息接收者（王校长向学生致欢迎词）；③转移活动中的转移物(客体)的接受者（政府为志愿者颁发证书）；④有意志或无意志主体行为所产生明显利益的接受者（Beneficiary）（老师指导学生做作业；我经常协助老师批改作业；④年青人给孕妇让座；）
伴体	同事	事件中的伴随者、排除者或作为基准的邻体：①事件中的伴随者(我陪妈妈去买衣服)；②事件中的排除者(除了英语之外我还会日语)；③主体比较或测量的基准(姐姐比妹妹大五岁)；④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)；⑤与主体相当但与主体属于不同类的客体(每节约 10 升汽油就相当于减排 23 千克二氧化碳)；
	用事	事件中所用到的方法、时耗、工具以及材料等。按照客体用事种类的不同，用事一般可被分为：①方法类用事（王同学采用速记法记录了老师的上课内容）；②时耗类用事（我花了四个月终于写完了毕业论文）；③工具类用事（盗墓贼利用洛阳铲盗墓）；④材料类用事（施工队使用新型油漆装修卧室）等。
系体	属事	领属关系[20,22]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)等；
	类事	事件中存在的类同者，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的类同关系[20,22]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鳜鱼)；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份子)；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)；
	连事	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(这栋别墅靠近湖边)；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)；
情由	缘由	引起事件的原因或事件依照的根据：①引起事件的原因(老师喜欢学生诚实)；②事件所依照的根据(法庭依照法律判决)；
	意图	事件所要达到的目的，主要包括：①事件主体的意图(她怂恿丈夫争夺家族生意)；②事件客体的意图(我经常协助老师批改作业)；
时空	时间	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)；
	位事	事件所在的物理或信息空间(①我在中科院计算所学习；②张经理在微博发表声明)；
	空间+	事件所在的社会空间：①小明在逆境中拼搏；②杨过于困境中领悟了新的剑法；③王萌当着记者的面揭露了事情的真相；
	范围	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(计算机方面我很擅长)；
描述	属性+	事件中对象的某一个方面的信息，通常与值事搭配使用：华为 mate30 手机的官方售价是 4999 元；
	性质+	事件中对象本身所具有的与其他对象不同的特征：这种新型金属具有良好的导电性。需要注意的是，性质和属性很容易被混肴，但其实它们是不同的，性质通常可以被写成：性质(对象)。
度量	物量	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	事件本身持续的时段或迄今所经历的时段。①事件本身所持续经历的时段(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(这袋薯片已经过期了半年)；
	频量	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事件中实体行动重复的次数(我一年去北京四趟)；②事件中事物变化重复的次数(蛇大约每年可以蜕皮 3～4 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)；
	值事+	事件中一个对象在某个属性上的取值，通常与属性搭配使用：华为 mate30 手机的官方售价是 4999 元；

2.2.2 辅助周边语义角色

辅助周边语义角色表示在事件中损失利益（受损者）或者获得收益（受益者）的语义角色。事实上，在实际语义角色标注过程中，我们经常遇到一些句子成分在语句中损失利益或获得收益。对于这些句子成分，除了需要标注必要的主要周边语义角色外，我们还应该标注相关的辅助周边语义角色，即受益者（Benefactive）或者受损者（Malefactive）。例如，在“法轮功伤害了无数的家庭”这个句子中，句子成分“无数的家庭”被标注为受事，但是它还扮演着“受损者”的角色。通过结合事件中的主要周边语义角色和辅助周边语义角色，我们可以更直观地得到事件的语义关系。

3 一种半自动的汉语语义角色数据集构建方法

在本节中，我们基于第 2 节中改进的汉语语义角色分类体系，进行汉语语义角色数据集的构建。基于细粒度的语义分析和知识获取等任务的需要[12-15]，在参考了现有的汉语语义角色的标注方法后，我们提出了一套详细的细粒度的汉语语义角色的标注流程，如图 1 所示。在从汉语语料库中提取出相关汉语语料后，本文提出的汉语语义角色的标注流程主要分为四个环节：（1）中枢语义角色和主要周边语义角色标注；（2）辅助周边语义角色标注；（3）语义角色的双重标注；（4）不确定语义事件的语义角色标注。

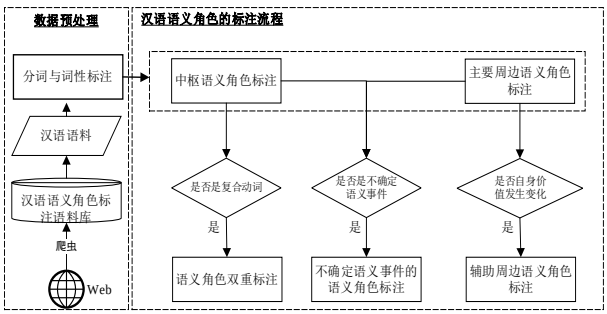

图 1 汉语语义角色的标注流程

3.1 语料收集与整理

我们课题组经过周丹[12]、臧良俊[13]、王亚[14]以及方芳[15]等同学的科研工作，积累了一个良好的汉语语义角色标注语料库，用于研究和测试新的语义角色标注和知识获取算法。该汉语语义角色标注语料库中的语料来源于商业网站、门户网站、新闻网站等，涉及如汽车[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。我们用这个汉语语义角色标注语料库作为本文汉语语义角色标注数据集构建

[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/

的语料资源。

现有的大部分汉语语义角色标注数据集在语义角色标注方面是粗粒度的，这造成了其在后续的语义分析、知识获取和自动问答等应用中存在缺陷。例如，在 Chinese Proposition Bank 中标注的一条语句“河北省/NR/B-ARG0 科学院/NN/I-ARG0 微生物/NN/I-ARG0 研究所/NN/I-ARG0 所长/NN/I-ARG0 刘荣昌/NR/I-ARG0 研究员/NN/E-ARG0 ，/PU/O 从/P/O 一九八四年/NT/O 开始/VV/O 与/P/O 科研/NN/O 人员/NN/O 苦心/AD/O 钻研/VV/O ，/PU/O 筛选/VV/O 出/VV/O 两/CD/O 株/M/O 分解/VV/O 土壤/NN/O 矿物/NN/O 钾/NN/O 能力/NN/O 很/AD/O 强/VA/O 的/DEC/O 硅酸盐/NN/O 菌株/NN/O ，/PU/O 采用/VV/O 当代/NN/O 最新/JJ/O 发酵/NN/O 工艺/NN/O ，/PU/O 于/P/B-ARGM-TMP 一九八八年/NT/E-ARGM-TMP 在/P/B-ARGM-LOC 国内/NN/E-ARGM-LOC 率先/AD/S-ARGM-ADV 生产/VV/rel 出/VV/O 首/OD/B-ARG1 批/M/I-ARG1 硅酸盐/NN/I-ARG1 菌剂/NN/I-ARG1 生物/NN/I-ARG1 钾肥/NN/E-ARG1 。/PU/O”中，存在“筛选”、“采用”、“生产”等多个中枢语义角色，但只标注了“生产”这一中枢语义角色，这会丢失很多语义信息。例如，在问答系统中问计算机“刘荣昌采用了什么工艺生产了硅酸盐菌剂生物钾肥？”，由于语义角色标注的不完整，计算机无法回答这个问题。考虑到较长的语句不便于细粒度的语义分析与标注，因此在本文语义角色标注语料库建设初期的语料筛选方面，与现有的语义角色数据集所用的语料相比，我们选取的语句绝大部分只有一个中枢语义角色且其句子的平均长度较短。对于较长的句子，我们会通过句中逗号、分号和句号等标点符号进行截断，如果截断后的分句仍然具备完整的语义信息，我们将选取该截断后的分句进行语义角色标注。尽管我们选取语句的平均长度短于现有的其他语义角色数据集中的语句，但我们标注的语句蕴含更细节的语义信息，这不仅体现在我们的主要周边语义角色种类更丰富，还体现在我们增加了辅助周边语义角色标注、语义角色的双重标注以及不确定语义事件的语义角色标注的标注方式，能够为标注的语句挖掘更多的语义信息，这是现有其他语义角色数据集不具备的优势，也是本文的创新之处。在未来的工作中，我们会考虑长语料的细粒度语义角色标注。

3.2 基于语义分类与描述框架的中枢和主要周边语义角色标注方法

为了满足各个领域 Web 文本语义分析与知识获取方法的研究需求，我们课题组构建了一种基于语义分类和描述框架 FSTD 的知识库[13,22]。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、常量终结符组成的序列，其中，语义非终结符是指形为“<主要周边语义角色：类型>”的非终结符，“类型”是充当主要周边语义角色的对象的类型。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>表示的是“买”事件的一个语义文法。“<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。其中，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”为对主要周边语义角色类型所做的语义限制，主要周边语义角色是语义层面上的定义，而对象类型是知识层面上的定义。

基于 FSTD 的语义文法，方芳[15]实现了一个鲁棒 Earley 语义解析器，能够对给定的中文文本进行解析。鲁棒 Earley 语义解析器是基于 Earley 算法进行改进而产生的语义解析器，加入了容错解析等功能，用于判断一个中文文本是否可以被 TSTD 中的语义文法所识别，并生成识别后的语义解析树。目前语义分类与描述框架 FSTD 的语义文法在公开的 Web 数据上覆盖率已经达到 71.4%[15]，即能解析语料库中的大部分句子，对于无法解析的句子，则通知 FSTD 维护人员进行语义文法的扩充。

考虑到中枢语义角色和主要周边语义角色是 FSTD 语义文法中的重要组成部分，本文充分利用 FSTD 中的语义文法和鲁棒 Earley 语义解析器对给定中文文本进行中枢和主要周边语义角色的自动标注，具体标注过程如下：

（1）首先对待标注的汉语语句用 FoolNLTK[5] 进行分词和词性标注。例如，语句“胡老师祝愿大家新年快乐”的分词结果为“胡/nr1 老师/n 祝愿/v 大家/rr 新年/t 快乐/an”。

（2）利用鲁棒 Earley 语义解析器对分词后的语句进行自动解析，得到语义解析树。图 2 给出了“胡老师祝愿大家新年快乐”的最优语义解析树。

```
<Start>
|---|-<祝福类语句>:
|---|---|<施事：认知主体>:
|---|---|---|<名词短语>:
|---|---|---|---|<基本名词短语>:
|---|---|---|---|---|/nr1：胡
|---|---|---|<名词短语>:
|---|---|---|---|<基本名词短语>:
|---|---|---|---|---|/n：老师
|---|---|<中枢语义角色>:
|---|---|---|祝愿/：祝愿
|---|---|<向事：认知主体>:
|---|---|---|<名词短语>:
|---|---|---|---|<基本名词短语>:
|---|---|---|---|---|/r：大家
|---|---|<客事：信息>:
|---|---|---|<字符串变量>：新年快乐
|---$/：$
```

图 2 FSTD 语义文法解析得出的最优解析树

[5] https://github.com/rockyzhengwu/FoolNLTK

（3）从最优语义解析树中获得语句中相关成分的语义角色信息，产生一个初步的语义角色标注框架，参见图 3。

```
defframe 胡老师祝愿大家新年快乐
{
    分词结果：胡/nr1 老师/n 祝愿/v 大家/rr 新年/t 快乐/an
    中枢语义角色：祝愿
    主要周边语义角色标注：
        施事：胡老师      施事类型：认知主体
        向事：大家        向事类型：人群
        客事：新年快乐    客事类型：信息
}
```

图 3 从最优解析树中产生的语义角色标注框架

（4）由于目前鲁棒 Earley 语义解析器得到的最优语义解析树正确率只有 58%左右，因此，我们需要对图 3 中得到的初步语义角色标注框架进行人工审核。通过人工修改和补充，我们获得关于语句“胡老师祝愿大家新年快乐”的语义角色标注结果，如图 4 所示，主要包含 5 个部分，分别为：①输入部分 defframe 表示待标注的原始语句；②分词结果表示的是对原始语句进行分词后的结果；③中枢语义角色标识句子成分“祝愿”是中枢语义角色；④主要周边语义角色标识句子成分对应的主要周边语义角色，例如，句子成分“胡老师”是施事等。此外，我们还保留了 FSTD 语义文法中主要周边语义角色的对象的类型，如充当施事“胡老师”的类型是认知主体等。⑤为了便于后期对标注语句的审查和理解，我们手动地在标注结果中增加“标注依据”，表示标注这些主要周边语义角色所参考的标准。

```
defframe 胡老师祝愿大家新年快乐
{
    分词结果：胡/nr1 老师/n 祝愿/v 大家/rr 新年/t 快乐/an
    中枢语义角色：祝愿
    主要周边语义角色标注：
        施事：胡老师      施事类型：认知主体
        向事：大家        向事类型：人群
        客事：新年快乐    客事类型：信息
    标注依据：客事，表示主体传递的信息(客事判断标准③)，作为向
              事的认知主体是信息的接收者(向事判断标准②)。
    标注时间：2019.11.05        标注人：宋衡
}
```

图 4　从最优解析树中得到的标注结果

基于 FSTD 语义文法的语义角色标注方法只能标注句子中的中枢和主要周边语义角色，而对于辅助周边语义角色、语义角色的双重标注以及不确定语义事件的语义角色标注这类语义性更强的标注需要采用人工的方法进行手动标注。

3.3 辅助周边语义角色标注

如第 2 节所述，在实际语义角色标注过程中，经常会遇到句子成分在事件中损失利益或者获得收益，为了获得这部分语义信息，我们必须为其标注相关辅助周边语义角色，即受益者或者受损者，下面，我们进行详细说明。

定义 1（受益者，Benefactive）：受益者表示的是在事件中枢语义角色动作发生之后，自身价值取向增加的周边语义角色。具体地，在事件 E 中，存在周边语义角色 PSRs 和中枢语义角色 PSR，如果 PSRs 初始自身价值为V_1，在中枢语义角色 PSR 动作发生之后，PSRs 自身价值发生了变化，其值的大小变成了V_2，且$V_2 > V_1$，则周边语义角色 PSRs 在辅助周边语义角色中被标注为受益者。

定义 2（受损者，Malefactive）：受损者表示的是在事件中枢语义角色动作发生之后，自身价值取向减少的周边语义角色。具体地，在事件 E 中，存在周边语义角色 PSRs 和中枢语义角色 PSR，如果 PSRs 初始自身价值为V_1，而在中枢语义角色 PSR 动作发生之后，PSRs 自身价值发生了变化，其值的大小变成了V_2，且$V_2 < V_1$，则周边语义角色 PSRs 在辅助周边语义角色中被标注为受损者。

基于定义 1 和 2，如果一个周边语义角色在事件中枢语义角色动作发生之后，自身价值取向增加，则我们将其标注为辅助周边语义角色“受益者”。反之，我们则将其标注为辅助周边语义角色“受损者”。我们在图 4 的标注结果上进行辅助周边语义角色的手工标注，如图 5 所示，句子成分“大家”是受益者。

```
defframe 胡老师祝愿大家新年快乐
{
    分词结果：胡/nr1 老师/n 祝愿/v 大家/rr 新年/t 快乐/an
    中枢语义角色：祝愿
    主要周边语义角色标注：
        施事：胡老师      施事类型：认知主体
        向事：大家        向事类型：人群
        客事：新年快乐    客事类型：信息
    辅助周边语义角色标注：
        受益者：大家
    标注依据：客事，表示主体传递的信息(客事判断标准③)，作为向事
              的认知主体是信息的接收者(向事判断标准②)。
    标注时间：2019.11.05        标注人：宋衡
}
```

图 5　辅助周边语义角色“受益者”的标注

辅助周边语义角色的标注非常重要，通过与主要周边语义角色的配合，我们可以更准确地推理出给定文本所要表达的语义信息。例如，在图 5 中，我们知道文本“胡老师祝愿大家新年快乐”一共被标注了施事、向事与客事，同时向事对应的句子成分“大家”被标注了受益者，那么从这 3 个主要周边语义角色和 1 个辅助周边语义角色，我们可以推导出给定文本蕴含着“施事传达了良好的信息给向事”、“施事对向事是友好的”等更详细的语义信息。

在对语料的标注过程中，我们发现标注人员在标注辅助周边语义角色“受益者”和“受损者”存在立场问题。下面我们以一个例句进行说明。

例句 4：红军攻占了阳新县城

例句 4 的中枢语义角色是“攻占”，两个主要周边语义角色“红军”和“阳新县城”分别是施事和是受事。然而，在对“阳新县城”进行辅助周边语义角色的判别时，我们课题组的标注人员产生了分歧：部分标注人员认为“阳新县城”是“受损者”，因为“阳新县城”是被攻击方，被攻击表示自身价值受损。另一部分标注人员觉得“阳新县城”是“受益者”，他们给出的理由是“红军”

是正义的一方，攻打阳新县城能够解放阳新县城的人民，摆脱阳新县城被统治者统治的状态。事实上，标注人员产生分歧的根本缘由是他们所处的立场不同，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，阳新县城被攻击会动摇统治者的地位。而认为“阳新县城”是“受益者”的标注人员所处的立场是红军和阳新县城百姓的立场，阳新县城被攻击能够解放阳新县城的百姓，红军能够给他们更好的生活。在标注辅助周边语义角色时，对于这类因为标注立场不同而存在分歧的语料，我们采用绩效风险的方法进行处理，即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，尽量减少标注。

3.4 语义角色的双重标注

语义角色的双重标注是基于我们对课题组语料审查分析过程中发现的一个特殊语义现象，它能够为文本语义理解挖掘更多的语义信息。语义角色的双重标注意味着在为事件角色标注了一个主要周边语义角色后，还需要为这个事件角色添加另一个主要周边语义角色。我们以引言中的例句 3“八路军主力已退守沂蒙山”进行说明。

例句 3 的中枢语义角色是“退守”，它有两个主要周边语义角色，分别为“八路军主力”和“沂蒙山”。“八路军主力”是“退守”动作的发起者，因此“八路军主力”是施事，然而，由于“退守”是一个复合动词，它的释义为“后退并防守”，其语义重心更偏向于“防守”，因此该事件对应的客体“沂蒙山”为“受事”。这样的标注方法可以让计算机理解“沂蒙山”是“八路军主力”防守并保护的对象，然而，在问答系统中，如果计算机被问“八路军主力现在所处的位置在哪？”，由于“沂蒙山”被标注成为“受事”，计算机就无法回答这个问题。这是由于在理解“退守”时，我们理解其语义更偏向于“防守”造成的，但是“退守”还有另一层“后退”的意思，它解释了八路军现在所处的位置。因此，为了解决这个问题，我们提出了语义角色双重标注的语义角色标注方法，即沂蒙山不仅被标注为受事（第一个标注的主要周边语义角色），还被标记为宿事（第二个标注的主要周边语义角色），如图 6 所示。

```
defframe 八路军已退守沂蒙山
{
      分词结果: 八路军/n 主力/n 已/d 退守/v 沂蒙山/ns
      中枢语义角色: 退守
      主要周边语义角色标注:
          施事: 八路军    施事类型: 认知主体
          受事: 沂蒙山    受事类型: 物体
          宿事: 沂蒙山    宿事类型: 地区
      判断依据: 受事，表示主体守护的客体(受事判断标准⑤)；宿事，主
                体经过活动后所在的地点(宿事判断标准①)。
      标注时间: 2019.12.27        标注人: 宋衡
}
```

图 6 主要周边语义角色的双重标注

图 6 中主要周边语义角色的双重标注方式能够为中文文本语义的理解带来了更多的语义信息，我们不仅知道沂蒙山是“八路军主力”防守的对象，还知道八路军主力正处于沂蒙山。需要注意的是，不是所有的语句都具有需要双重标注的语义角色，这类语句一般具有一些特殊的中枢语义角色，在我们对课题组语料库审查的过程中，我们发现这些特殊的中枢语义角色通常以复合动词的形式存在。

3.5 不确定语义事件的语义角色标注

不确定语义事件指存在歧义的事件，一般需要结合上下文消除歧义。然而，就单个事件而言，一个事件中被标注的语义角色是独立且与上下文中的其他事件的语义角色无关，这是不正常的。在语义角色标注过程中，我们发现造成不确定语义事件发生的原因主要有两个：（1）中枢语义角色造成事件语义不确定性；（2）周边语义角色造成事件语义不确定性。下面，我们分别进行介绍。

（1）中枢语义角色造成的事件语义不确定性：指因中枢语义角色多义性造成其所在事件存在语义不确定性。下面我们给出一个例句进行说明。

例句 5：康诺利拜访了教父皮尔斯

在例句 5 中，它的中枢语义角色是“拜访”。然而，在现代汉语中，“拜访”有两个基本释义：（1）短时间看望（到长辈或亲友等处问候）；（2）指敬词，表示看望并谈话。因此，要完成例句 5 中语义角色标注，我们必须结合上下文。下面，我们为例句 5 补充上下文，形成事链[20]例句 6 和例句 7：

例句 6: 教父皮尔斯很有人缘, 有一天路过他家时, 康诺利拜访了教父皮尔斯, 教父皮尔斯很开心。

例句 7: 康诺利遇到了一些困惑, 康诺利拜访了教父皮尔斯, 教父皮尔斯的一席话让康诺利茅塞顿开。

```
defframe 康诺利拜访教父皮尔斯
{
      分词结果: 康诺利/nrf 拜访/v 了/ule 教父/n 皮尔斯/nrf
      中枢语义角色: 拜访
      {
          主要周边语义角色标注:
              施事: 康诺利              施事类型: 认知主体
              向事: 教父皮尔斯          向事类型: 认知主体
          辅助周边语义角色标注:
              受益者: 教父皮尔斯
      }
      {
          主要周边语义角色标注:
              施事: 康诺利              施事类型: 认知主体
              源事: 教父皮尔斯          源事类型: 认知主体
          辅助周边语义角色标注:
              受益者: 康诺利
      }
      判断依据: 句子成分“拜访”作中枢语义角色，具有二义性: ①如
                果理解为短时间看望（到长辈或亲友等处问候），则句
                中对应的客体成分为向事（向事的判断标准②③）；②
                如果理解为指敬词，表示看望并谈话，则句中对应的客
                体成分为源事（源事的判断标准④⑤）；
      标注时间: 2020.01.02        标注人: 宋衡
}
```

图 7 中枢语义角色多义性的语义角色标注

对于例句 6 和例句 7，我们可以知道，例句 6 中“拜访”是基本释义中的“短时间看望（到长辈或亲友等处问候）”，例句 7“拜访”是基本释义中的“指敬词，表示看望并谈话”。因此，对于这类拥

有二义性中枢语义角色的语句，且没有足够上下文信息支撑标注人员确定事件代表的语义信息，我们需要标注可能的主要周边语义角色以便表示句子完整的语义信息，如图 7 所示。按照对例句 6 和例句 7 的理解，我们进行了两种不同的语义角色标注方式：①如果“拜访”被理解为“短时间看望（到长辈或亲友等处问候）”，则“康诺利”被标注为施事，而“教父皮尔斯”被标注为向事以及受益者；②如果“拜访”被理解为“指敬词，表示看望并谈话”，则“康诺利”同时被标注为施事以及受益者，而“教父皮尔斯”被标注为源事。

```
defframe 熊猫破坏了系统硬盘数据
{
      分词结果：熊猫/n 破坏/v 了/ule 系统/n 硬盘/n 数据/n
      中枢语义角色：破坏
      {
          主要周边语义角色标注：
              施事：熊猫              施事类型：认知主体
              变事：系统硬盘数据      变事类型：实体
      }
      {
          主要周边语义角色标注：
              施事：代施事            施事类型：实体
              变事：系统硬盘数据      变事类型：实体
      }
       判断依据：句子成分“熊猫”作周边语义角色，具有二义性：①
                 如果理解为哺乳动物熊猫，则句中对应的主体成分为
                 施事（施事的判断标准①）；②如果理解为计算机蠕虫
                 病毒熊猫，则句中对应的主体成分为代施事（代施事
                 的判断标准③）；
       标注时间：2020.01.16        标注人：宋衡
}
```

图 8　周边语义角色多义性的语义角色标注

（2）周边语义角色造成的事件语义不确定性：指的是因为周边语义角色存在多义性造成了其所在事件存在语义不确定性。对于这种情况，需要结合上下文并进行词义消歧才能准确地完成语义角色的标注。下面我们给出一个例句进行说明。

例句 8：熊猫破坏了系统硬盘数据

在例句 8 中，句中的成分“熊猫”存在多语义性，因为这个词既可以指代哺乳动物熊猫，也可以指代计算机蠕虫病毒熊猫。如果句中的成分“熊猫”指的是哺乳动物，那么其对应的语义角色是施事（施事的判断标准①）。如果句中的成分“熊猫”指的是计算机蠕虫病毒，那么其对应的语义角色是代施事（代施事的判断标准③）。因此，对于这类事件语料，我们参照图 7 的标注方式，增加另一种标注以便表示其完整的语义信息，并在注释中进行说明，如图 8 所示。

3. 6 语义角色标注后的审查

在我们进行语义角色标注的过程中，我们发现语义角色在语料中的出现和分布存在一定的规律，我们将这种规律称为语义角色约束。语义角色约束可以帮助我们设计相关算法对语料库中标注的语料自动地进行初步地审查，减轻后期人工复审的压力。目前，我们发现 3 种语义角色约束关系，分别为主体客体数量约束、主要周边语义角色搭配约束、中枢语义角色模式一致性约束。为了能够清晰地表示这三种语义角色约束关系，我们以节点作为语义角色，以边作为约束关系，绘制了语义关系约束图，如图 9 所示。下面我们介绍这三种约束关系。

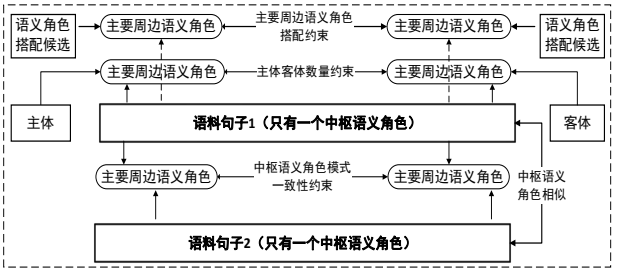

图 9　主要周边语义角色关系约束图

（1）主体客体数量约束：指在只有一个中枢语义角色的语句中，主体或者客体中的主要周边语义角色出现的个数必须均低于 2。例如，在只有一个中枢语义角色语句中，施事和雇施事作为主体不能同时出现，受事和客事作为客体不能同时出现等。

（2）主要周边语义角色搭配约束：在我们的语义角色标注体系中，有些语义角色是搭配使用的，这些搭配使用的语义角色在只有一个中枢语义角色的语句中会同时出现。在我们对语料进行语义角色标注的过程中，我们发现常见搭配使用的语义角色包括{领事⇌属事}[6]、{属性⇌值事}、{感事→客事}、{当事←连事}、{(施事，致事) →意图}、{施事|代施事|当事←(客事，源事)}、{施事|代施事←(受事，源事)}、{施事|代施事←(客事|受事，向事)}等。例如，在只有一个中枢语义角色的语句中，如果出现了领事但没有出现属事，则该语句的语义角色可能标注错误，反之亦然。

（3）中枢语义角色模式一致性约束：指对于中枢语义角色语义相近的语句，其标注的主要周边语义角色类型应该是相同的，即两个中枢语义角色意思相近的语句共享同样的主要周边语义角色类型。例如，在以“祝福”为中枢语义角色的语句中，其对应的主要周边语义角色有施事、向事和客事，那么在以“祝愿”为中枢语义角色的语句中，其对应的主要周边语义角色也应该是施事、向事和客事。

上述语义角色约束关系只能初步地找出可能标注出错的语料，并不能最终确定是否真的标注出错或者是句子中的哪个成分是错误的标注。因此，接下来需要通过人工的方法进行最终标注结果的确定，即人工复审确认。人工复审确认主要是为了提高语义角色标注的一致性，复审人员由另一个标注人员担任，其在复审过程中被要求对复审的句子重新进行标注。如果复审人员标注的每个句子成分对应的语义角色标签与初次标注结果一致，则将该

[6] $\{a \rightharpoonup b\}$表示有语义角色 a 必定存在 b；$\{a \leftharpoondown b\}$表示有语义角色 b 必定存在 a；$\{a \rightleftharpoons b\}$表示语义角色 a 和语义角色必然同时存在；

句确认标注完成，否则让课题组的权威老师进行判定。我们用 Kappa 系数[23]衡量语义角色标签标注的一致性，通过对数据标注的结果进行统计分析，初次标注人员和复审人员标注相同语义角色标签的 Kappa 系数为 0.8196，表明人工标注具有较好的一致性，达到了我们预期的目标。

4 基准实验

基于第 3 节的汉语语义角色数据集构建方法，截至到撰写本论文时，我们一共完成了 9550 条汉语语句的语义角色标注，初步形成了一个细粒度的语义角色标注数据集。在这个数据集中，目前一共拥有 9423 个中枢语义角色，29142 个主要周边语义角色，3745 个辅助周边语义角色、172 条语句被进行了语义角色的双重标注以及 104 条语句被进行了不确定语义事件的语义角色标注。

目前，在语义角色数据集的研究应用方面，主要周边语义角色的自动识别被广泛研究。为了测试我们标注语料的合理性以及构建的语义角色数据集的有效性，我们用本文的语义角色数据集进行了关于主要周边语义角色自动识别的基准实验。具体实验过程如下：

我们将主要周边语义角色的自动识别问题看作是序列标注问题。对于上述标注完成的语义角色数据集，我们采用 BIOES 的方式进行转换，其中 B 表示开始，I 表示中间，E 表示结尾，S 表示单个分词词语，O 表示其他，用于标记与中枢语义角色和主要周边语义角色无关的分词词语。我们将中枢语义角色对应的句子成分标注为 Pivot。此外，对于语义角色双重标注和不确定语义事件的语义角色标注的语料，我们选取其第一种标注方式进行 BIOES 转换。对于图 5 中的例句，最后转换得到的 BIOES 的格式如图 10 所示：

```
sentence：胡/nr1 老师/n 祝愿/v 大家/rr 新年/t 快乐/an $/
PPSR：  [  施事  ] [Pivot] [向事] [  客事  ]
BIOES：B-施事 E-施事 Pivot S-向事 B-客事 E-客事
```

图 10 主要周边语义角色对应的 BIOES 格式

基于上述转化后 BIOES 格式的主要周边语义角色数据集，我们采用了一个经典的序列标注模型 Bi-LSTM+CRF[7]作为基线模型进行主要周边角色的自动识别工作。考虑到在现有的中文语义角色数据集中，Chinese Proposition Bank 1.0(CPB 1.0)是公开的且被广泛地应用于中文语义角色自动识别的研究中，我们用这个基线模型将本文的语义角色数据集与 CPB 1.0 进行了比较。表 2 展示的是本文语义角色数据集和 CPB 1.0 的构建差异对比。

表 2 本文语义角色数据集和 CPB 1.0 的构建差异对比

数据集	CPB 1.0	本文语义角色数据集
规模	37183	9550
语义角色标签种类数量	19	32
语义角色标签密度	0.25	0.67
辅助周边语义角色标注	部分(受益者作为附加语义角色，定义模糊)	有
语义角色双重标注	无	有
语料来源领域	新闻	新闻、汽车、数码、百科知识

从表 2 中可以看出，本文语义角色数据集目前在标注规模上还小于 CPB 1.0，但在语义角色标签种类数量和语义角色标签密度（语义角色标签数/标签总数，数值越小代表语义角色数据集中的 O 标签越多）都大于 CPB 1.0，这表明本文的语义角色数据集拥有比 CPB 1.0 更细粒度的语义信息。由于目前我们的语义角色数据集只有 9550 条语句，为了更好地体现出比较效果，我们从 CPB 1.0 中随机选取了 9550 条语句，保证其数量与我们的数据集数量相同。我们将我们构建的语义角色数据集和 CPB 1.0 中的 9550 条语句随机打乱，并按照 7：2：1 的比例划分为训练集和验证集和测试集，最后得到拥有 6685 条语句的训练集，拥有 1910 条语句的验证集以及拥有 955 条语句的测试集。我们采用准确率(Precision, P)、召回率(Recall, R)以及 F1 值来衡量最终的定量实验效果。Bi-LSTM+CRF 的基线模型在两个数据集的定量实验效果如表 3 所示。

表 3 Bi-LSTM+CRF 的基线模型在两个数据集的定量实验效果

数据集	预训练模型	验证集			测试集		
		P	R	F1	P	R	F1
CPB 1.0	无	68.62%	68.62%	68.62%	67.59%	67.76%	67.67%
	BERT	75.95%	76.11%	76.03%	75.05%	75.29%	75.17%
本文语义角色数据集	无	72.84%	73.81%	73.32%	71.94%	72.68%	72.31%
	BERT	79.19%	79.88%	79.53%	78.37%	78.97%	78.67%

从表 3 中，我们可以看出无论对于 CPB 1.0 还是本文语义角色数据集，基于 BERT[8]的预训练模型

[7] https://github.com/scofield7419/sequence-labeling-BiLSTM-CRF

[8] https://github.com/google-research/bert#pre-trained-models

均能够很明显地提高主要周边语义角色的识别准确率和召回率。此外，我们还注意到采用同样的 Bi-LSTM+CRF 基线模型和预训练模型，本文的语义角色数据集在包括准确率、召回率以及 F1 值三个方面均比 CPB 1.0 更好，但这并不能说明本文语义角色数据集的语义角色识别任务比 CPB 1.0 容易，出现表 3 中的实验效果是因为本文语义角色数据集在初步构建时所选用语料的长度比 CPB 1.0 短，而 Bi-LSTM+CRF 基线模型更容易捕捉短句中词与词之间的上下文依赖关系。事实上，本文构建的语义角色数据集在主要周边语义角色识别方面比 CPB 1.0 更具挑战性，这是因为本文语义角色数据集的主要周边语义角色的种类比 CPB 1.0 更细致，蕴含的语义信息也更细腻，这增加了主要周边语义角色自动识别的难度。

除了表 3 中的定量分析，我们还对 Bi-LSTM+CRF 基线模型在本文语义角色数据集的实验结果进行了定性错误分析。我们初步将错误类型分为两种：

（1）与常识无关的错误。例如，实验结果中存在将“李嬷嬷正在斥骂着宫人们”识别为“李/B-施事 嬷嬷/B-代施事 正在/O 斥骂/Pivot 着/O 宫/B-受事 人们/E-受事”，其错误在于施事对应句子成分后接上了代施事，辱骂对应客体句子成分是向事，而不是受事。对于这种与常识无关的错误，我们可以通过增大语义角色数据集的规模，设计更先进的神经网络模型以及更具有针对性的损失函数等方法来解决。

（2）与常识有关的错误。这种类型的错误很难甚至无法通过基于神经网络统计模型的方法来解决，我们总结了实验结果中四个典型容易识别错位的案例，如表 4 所示。

表 4 因常识问题容易混肴出错的四个典型案例

序号	案例
1	①他（施事）抢劫了银行（源事） ②他（施事）抢劫了珠宝（受事）
2	①特鲁希略家族（施事）大量掠夺土地（受事） ②特鲁希略家族（施事）掠夺了欧洲大陆（源事）
3	①熊猫贝贝（施事）破坏了硬盘数据（变事） ②熊猫病毒（代施事）破坏了硬盘数据（变事）
4	①王老师（施事）回复了这个问题（客事） ②王老师（施事）回复了这个人（向事）

造成表 4 中与常识有关的语义角色识别错误发生的根本原因在于作为一种浅层的语义分析分析方法，语义角色标注在知识和常识的表达方面存在先天的缺陷，以“他抢劫了银行”和“他抢劫了珠宝”这两个句子为例，Bi-LSTM+CRF 基线模型会将句子成分“珠宝”和“银行”均识别为受事，但在我们的语义角色标注体系中，“银行”应该被标注为源事，“珠宝”应该被标注为受事，这种标注方式在我们人脑中蕴含的一个常识就是银行是一个金融机构，他抢劫了银行预示着他把银行中的贵重物品拿走了，珠宝是一种贵重物品，他抢劫了珠宝预示着他把珠宝拿走了。这种知识和常识信息对语义分析至关重要，能够很好地服务于后续关于语义方面的知识抽取和常识获取等下游任务。

为了能够形式化表示上述这种知识和常识信息，我们课题组在曹存根研究员的主持下正在研发一种全息事件网络（Holographic Event Network, HEN），其致力于为深层的语义分析、知识获取和常识获取等研究打下基础。HEN 是一种包含过程事件网络层和状态事件网络层的事件网络，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[9]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。

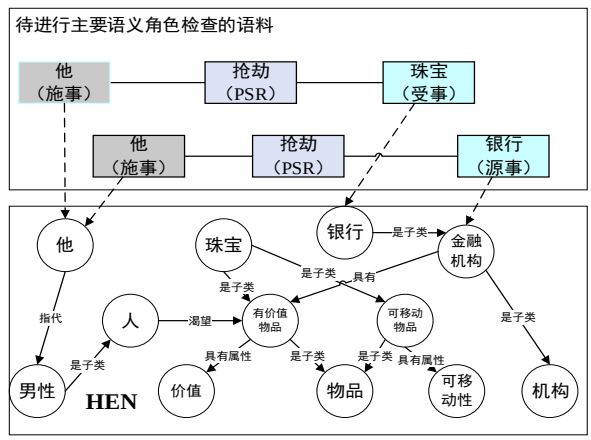

图 11 语义角色在 HNN 中被自动检查的过程

HEN 的建立为解决上述因常识问题造成主要周边语义角色识别出错的案例提供了思路。图 11 表示的是案例“他抢劫了银行”和“他抢劫了珠宝”到 HEN 状态事件网络层的映射。对于这种类型的错误，我们可以结合 HEN 并有针对性地利用规则的方式来进行自动检查，其具体算法如算法 1 所示。在算法 1 中，步骤 2-3 表示是对于如标注为 O 和 Pivot 等不是主要周边角色类型的句子成分，默认检查通过。步骤 5 表示的是将句子成分映射到 HNT 中的状态事件网络层。步骤 7-12 表示的是获取与句子成分在 HNT 中映射节点存在指代、是子类和同义词关系的节点，并判断这些节点的属性是否满足其所标识主要周边语义角色的判断标准，如果满足，则识别检查通过，反之则识别检查不通过。例如，句子成分“珠宝”是可转移物品和有价值物品的子

[9] https://github.com/commonsense/conceptnet5/wiki/Relations

类，其分别具有属性价值和属性可移动性，可移动性符合我们对受事的定义，因此，“珠宝”被标注为受事的检查通过。反之，句子成分“银行”是金融机构的子类，但是我们从 HNN 中只能知道金融机构具有有价值的物品，并不能得出金融机构具有属性可移动性，其不符合受事的定义，因此，“珠宝”被标注为受事的自动检查无法通过。

算法 1 基于 HNN 的主要语义角色识别检查算法

输入：$S=\{w_1/SR_1, w_1/SR_2, \dots, w_1/SR_n\}$，其中$w_i$表示句子中第$i$个句子成分，$SR_i$表示第$i$个句子成分被识别的主要周边语义角色类型；$PPSR=\{PPSR_1, PPSR_2, \dots, PPSR_n\}$，主要周边语义角色类型集合；$HNT\_S=<N, NE>$，其中$N=\{1,2,3,..,m\}$表示 HNT 状态事件网络层的节点集合，$NE_{jk}$表示节点$j$和节点$k$之间的关系；

输出：$Result=\{Re_1, Re_2, \dots, Re_n\}$，其中$Re_i$表示第$i$个句子成分是否识别正确，如果正确，则为 True，反之则为 False；

```
1、for all w_i/SR_i in S do
2、    if SR_i not in PPSR then
3、        Re_i ← True   ▷不是周边语义角色类型的句子成分，不需要检查
4、    else
5、        No ← HNT_S(w_i)
6、        Relation ← {指代, 是指类, 同义词}
7、        NoSet ← No ∪ Relation(No)  ▷得到Relation相关的节点
8、        for all Node in NoSet do
9、            if Node 具有的属性满足SR_i的判断标准 then
10、               Re_i ← True
11、       end
12         if Re_i == NULL then Re_i ← False
13、end
```

5 结束语

语义角色数据集的构建对自然语言语义分析和理解等研究有着重要的作用。本文深入研究了已有的语义角色分类体系，并对实际的汉语语料进行了详细的考察，提出了一种改进的汉语语义角色分类体系。在此基础上，以只有一个中枢语义角色的语料作为研究对象，提出了一种细粒度的汉语语义角色数据集构建方法。细粒度的语义信息不仅体现在我们主要周边语义角色种类的丰富性，还体现在我们语义角色标注步骤的多样性。最后，我们构建了一个拥有 9950 条语句的汉语语义角色数据集，并将其与公开的 Chinese Proposition Bank 语义角色数据集在一个 Bi-LSTM+CRF 的基线模型上进行了关于主要周边语义角色自动识别的实验对比。此外，我们还分析了 Bi-LSTM+CRF 基线模型在本文语义角色数据集识别错误的语句，并针对这些识别出错的语句初步提出了后期解决这些错误的思路。

目前，我们语义角色数据集的构建还有较长的路需要走，后期的工作重心将集中于扩大语义角色数据集的规模以及考虑如何更好地对多中枢语义角色的长语料进行细粒度的语义角色标注。此外，利用已有的语义角色数据集，设计相关语义角色的自动识别算法，提高语义角色识别的准确率和召回率，并将其运用于语义分析和知识获取等下游任务，也是我们未来的主要工作。

参考文献

[1] Kapetanios E, Tatar D, Sacarea C. Natural language processing: semantic aspects[M]. Florida: CRC Press, 2013.

[2] Che W, Li Z, Liu T. LTP: A Chinese language technology platform[C] //Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, 2010: 13-16.

[3] Abend O, Rappoport A. Universal conceptual cognitive annotation (UCCA)[C]//Proceedings of the 51st Meeting of the Association for Computational Linguistics, 2013: 228-238.

[4] 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.

[5] 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.

[6] Baker C F, Fillmore C J, Lowe J B. 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.

[7] Palmer M, Gildea D, Kingsbury P. The proposition bank: An annotated corpus of semantic roles[J]. Computational linguistics, 2005, 31(1): 71-106.

[8] Meyers A, Reeves R, Macleod C, et al. Annotating noun argument structure for NomBank[C]//Proceeding of the 2004 Language Resources and Evaluation Conference, 2004: 803-806.

[9] Xue N, Palmer M. Annotating the propositions in the Penn Chinese Treebank[C]//Proceeding of Sighan Workshop on Chinese Language Processing, 2003: 47-54.

[10] 李济洪, 王瑞波, 王蔚林, 等. 汉语框架语义角色的自动标注[J]. 软件学报, 2010, 21(4): 597-611.

[11] 袁毓林. 语义角色的精细等级及其在信息处理中的应用[J]. 中文信息学报, 2007, 21(4): 10-20.

[12] 周丹. 基于种子文法的汉语语义文法扩展方法研究[D]. 中国科学院大学硕士学位论文, 2015.

[13] 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.

[14] 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.

[15] 方芳. Web 文本语义分析与知识获取方法研究[D]. 中国科学院大学博士学位论文, 2019.

[16] Fillmore C J. The Case for Case[C]//Proceedings of the Texas Symposium on Language Universals, 1967: 13-15.

[17] 冯志伟. 从格语法到框架网络[J]. 解放军外国语学院学报, 2006, 29(003):3-11.

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

[19] 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.

[20] 鲁川. 知识工程语言学[M]. 北京: 清华大学出版社,

2010.
[21] 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.
[22] 王亚．基于语义分类的常识知识获取方法研究[D]．广西师范大学硕士学位论文, 2015.
[23] Carletta J. Assessing Agreement on Classification Tasks: The Kappa Statistic[J]. Computational Linguistics, 1996, 22(2): 249-254.

宋衡（1990—），博士研究生，主要研究领域为语义理解，知识获取。
E-mail：song_heng@foxmail.com

曹存根（1964—），通信作者，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

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