文章编号：1003-0077(2019)09-0001-11

融合概念与逻辑的中文深层语义描述体系

夏乔林[1,2]，穗志方[1,2]，常宝宝[1,2]，詹卫东[1,3]，张坤丽[4]，柯永红[1,2]

（1. 北京大学 计算语言学教育部重点实验室，北京 100871；
2. 北京大学 信息科学技术学院，北京 100871；
3. 北京大学 中文系，北京 100871；4. 郑州大学 信息工程学院，河南 郑州 450001）

摘　要：自然语言的语义理解涉及多个层面的问题，包括以谓词为中心的基本命题义、命题义之外的概念义、逻辑补足义等。目前主流的浅层语义分析主要集中在对命题义的分析上，缺少对概念义和逻辑义的支持，难以辅助计算机对文本的深度理解与推理。该文借鉴论元结构理论、事件语义学等相关语言学理论，突破语义角色标注等浅层语义分析的局限，建立了一种融合概念与逻辑的中文深层语义描述体系；并在该体系基础上，采用层层渲染的标注策略，构建了基于真实语料的大规模中文深层语义标注语料库，通过语言工程实践验证该描述体系的完备性和覆盖度。这一理论体系的建立和语言资源的构建，有望推动中文自动语义分析技术和人工智能等相关工作的创新发展。

关键词：中文语义；意义表示；资源构建

中图分类号：TP391　　**文献标识码**：A

Chinese Deep Semantic Representation with Concept and Logic

XIA Qiaolin[1,2]，SUI Zhifang[1,2]，CHANG Baobao[1,2]，ZHAN Weidong[1,3]，ZHANG Kunli[4]，KE Yonghong[1,2]

（1. MOE Key Laboratory of Computational Linguistics，Peking University，Beijing 100871，China；
2. School of Electronics Engineering and Computer Science，Peking University，Beijing 100871，China；
3. Department of Chinese Language and Literature，Peking University，Beijing 100871，China；
4. School of Information Engineering，Zhengzhou University，Zhengzhou，Henan 450001，China）

Abstract：The natural language understanding involves multiple categories of meaning，including propositions，modality，and temporal logic. The most popular study of shallow semantics is focused on the analysis of propositional meaning. Without supporting for conceptual meaning and deep logical meaning，it can be hardly used to assist the computer in deep understanding and reasoning of the text. Based on the theory of argument structures，event semantics，and construction grammar，this paper breaks through the limitations of shallow semantic analysis（e.g. semantic role labeling）and establishes a deep semantic representation system for concepts and logic. Based on a layered rendering annotation strategy，a large scale Chinese deep semantic annotated corpus is constructed，which also helps to verify the completeness and coverage of the description system by real practice. The establishment of this theoretical system and the construction of language resources are expected to promote the innovative development of Chinese automatic semantic analysis technology and artificial intelligence.

Keywords：Chinese semantics；meaning representation；resource construction

收稿日期：2018-11-26　　**定稿日期**：2019-01-31

基金项目：国家重点基础研究发展计划(2014CB340504)；国家自然科学基金(61751201)

0 引言

语义分析与理解在人工智能研究中的意义非比寻常,涉及语言学、计算机科学、机器学习,甚至认知科学等多个学科,是一个典型的多学科交叉研究课题。开展这项研究不仅对推动相关学科发展具有深远意义,同时也是揭示人脑理解语言的奥秘、实现真正人工智能的必经之路。

要理解自然语言,首先要理解自然语言所要表达的语义,尤其是句子的语义,因为句子通常是相对完整的自然语言基本意义表达单位。但是,什么表示形式才能够完整地描述句子的语义,这个问题一直困扰着研究者们,至今仍没有统一的答案。

要全面理解句子语义,涉及的因素非常多。理论上,一句话至少可以分解为五个层面的意义:基本命题义(句子的基本客观事件语义)、情态义(主观义,主要由句中的助动词表达)、事件关联义(多个动词表达的事件之间的关联含义)、构式义(简单加合其构成成分无法得出其整体意义的部分)、语用义(文化等因素带来的附加义或说话人的感情色彩义等)。如例1所示:

例1 小明考试没过,现在可能非常难过。

例1中就包含三个基本的命题:①小明+考试;②小明考试+(没)过;③小明+现在+难过。第一个命题中"小明"是"考试"的主体,第二个命题中"小明考试"是谓词性成分(VP)整体作主体,第三个命题"小明"是"难过"的主体,"现在"是"难过"发生的时间。这部分意义可以称为"谓词论元结构义",是命题义中的一种,是过去语言学家关注的重点,在许多著名的语料库中有所体现,例如,宾州大学命题树库(Penn PropBank)[1]、框架语义网(FrameNet)[2]等。此外,句中"没"代表对"过"的逻辑否定,而"非常"形容的是"难过"的程度,"难过"的时态是"现在时"。这些意义用于描述句子的主观和附加的含义,我们称为超命题义。另外,"过"表示"通过(考试)"而不是"经过"或者"超过"的含义,这涉及到更加底层的概念义的区分。这些基本命题义之外的意义显然也要包含在句子的整体语义解读中。

由例1可看出,自然语言语义的理解涉及多个层面的问题,只有把每个层面的问题都搞清楚,才有可能最终得到一个句子的完整语义解读。目前主流的浅层语义分析缺少对概念义和深层逻辑义的支持,难以辅助对文本内容的深度理解与知识推理任务。但就笔者所了解的情况来看,目前面向中文的深层语义描述体系研究在国内外几乎空白。本研究的目标就是从计算机计算和语言工程的角度,对文本蕴含的语义信息进行分层次、细粒度的深入挖掘,并借鉴论元结构理论、事件语义学、构式语法理论等现代汉语语法语义理论成果,突破语义角色标注等浅层语义分析的瓶颈,建立一套融合概念与逻辑的中文深层语义描述体系。在此基础上,我们还通过对大规模真实文本的标注,实现了对描述体系理论可行性的验证,并构建了一个能够服务于计算机辅助分析和理解的中文深层语义标注语料库。

1 相关工作

深层语义表达的目的是将整个句子转化为某种形式化表示,如基于依存的组合式语义表达式(dependency-based compositional semantic representation)[3]和谓词逻辑表达式(包括lambda演算表达式)。语义依存分析(broad-coverage semantic dependency parsing,SDP)项目及其语料库[4]即建立在依存理论基础上的描述体系之上。其提出的动机有两点,一是旨在将依存分析任务从树扩展到图,另一方面从语法扩展到语义,直接分析"谁对谁做什么"[4-6]。SDP包含两个步骤:依据语法建立依存结构,然后对所有的修饰词与中心词对(谓词论元结构)指定语义关系,展现形式主要是双词(Bilexical)语义依存图。相比于以PropBank和FrameNet为代表的浅层语义描述方式,SDP跨越了句子的表层句法结构而转向直接获取深层语义结构。但该描述体系也存在一些问题,比如一些常见语义现象被忽略,例如,否定及辖域、比较关系、所有格、从句连接等。SDP要求标注句中所有语义依存关系,也就是覆盖句中每个单词,但是没有触及单词组合而成的概念义,包括概念内部和概念作为整体与句中其他成分的关系。

哈尔滨工业大学与北京语言大学合作推出的H-SDP-v1语义依存表示体系[7],同样建立在依存理论基础上。车万翔等[8]整理后在SemEval-2012上组织了国际公开测评。在关系类型上,针对汉语句式特点定义了反关系和间接关系,分别用于描述动词修饰名词以及核心词是动词名词化形式两种情况。该语料库涵盖语义关系123种,但一些语义关系在语料中出现次数较少。此外句子全部来自新闻

语料，涵盖的语言现象可能受到一定限制。

另一方面，抽象语义表示（abstract meaning representation，AMR）[9] 是 Banarescu 等从融合多种语义资源角度出发提出的描述体系，动机是为将原本分离的多种描述体系包括命名实体、指代消解、浅层语义、篇章连接、时体等统一到一个逻辑表达形式中，即有根节点的有向语义图，图中每个边都有一个角色标注。它的一个显著特点是对文本所蕴含的语义进行了高度抽象，具体表现为：①将实词抽象为概念节点，动词和角色沿用 OntoNotes[10] 体系；②同一个 AMR 图可能表示各种各样语义相同的句子。这种抽象的表示确实使得语义的描述脱离了语法形态的限制，能够展示更深层次的语义关系，但这也导致最终句子的语义表示和句中单词不能一一对应，给之后自动分析算法的研究带来了困难，因为结构化表示与文本存在映射关系是很多成熟算法实施的先决条件，开发人工或自动文本对齐系统一方面会增加工程量，另一方面该系统产生的对齐错误将对最终分析的准确性产生影响。

抽象语义表示也可应用于中文（Chinese AMR，CAMR）[11]。CAMR 有向图的描述形式和英文 AMR 一致，标注规范为贴合中文句子特点对 AMR 语义关系进行了修改，忽略了一些难以标注的特例，如不标注“被”字句、“把”字句中的情态义等。并且为解决 AMR 本身无文本对齐的问题，提出将“单词—概念”对应关系纳入标注过程中。但 CAMR 公开的语料库规模目前只有 1 562 句，难以支持数据驱动的自动分析算法。

还有一种语义表达方式是一阶谓词逻辑表达式，典型的语料库有 GeoQuery[12]，训练集包含 880 个示例和一个有 800 个地理事实的数据库。每个示例包含一个提问（文本和对应的语义表达式）和一个回答，提问例如，“What are the major cities in Kansas?”，对应的语义表达式为“answer(C,(major(C),city(C),loc(C,S),equal(S,stateid(Kansas))))”. 基于这种语义表达方式的确能够帮助一些系统实现语义分析的终极目标，即自然语言的理解（对于提问）和推断（对于答案），但一个明显问题是：句型受到领域“美国地理”和数据量 880 句的限制，严重缺乏多样性，是一种领域定制的专用语料库，语种也仅限于英语。

总的来说，前述工作从一定程度上突破了句法树型结构和浅层语义分析的限制，但受到标注语料的领域和规模限制，并且从英文理论出发的表述形式，在应对中文文本时仍面临诸多问题。因此本文尝试借鉴论元结构理论[13]、事件语义学[14]、构式语法理论[15]，突破语义角色标注等浅层语义分析的局限，建立了一种从中文出发的融合概念与逻辑的深层语义描述体系，并在此基础上构建了基于真实语料的大规模中文深层语义标注语料库。语言工程实践也验证了该描述体系的完备性和覆盖度。

2　中文深层语义描述体系

2.1　中文语义的特点

对于给定句子，深层语义分析的目标是将整个句子转为某种形式化表示。这个过程涉及语言的多个层次和分面。目前中文语义角色标注、词义消歧、命名实体识别、指代消解、情感分析等研究都或多或少触及语义描述的不同侧面，但这些研究各自独立进行，拥有各自的研究目标，对应不同的评估策略和标注资源，表现出孤立局部的特点，且缺乏对情态义、时体义等高阶语义现象的计算处理。

除此以外，中文意合为主、缺乏形态标记的特点也为中文语义描述的研究带来困难。例如，同样是“要”字，在“领导要大家切实做好本职工作”中同一般动词“要求”的意思，可以标注为命题义的谓词，而“这个房间要干净一些”中的“要”是助动词，不是命题义的标注对象。情况相似的还有一些形式动词、虚化动词、泛义动词、谓词性结构中心词等。除了谓词选择问题，缺乏形态标记还给概念义消歧、论元成分确定、超命题义的识别都带来困难。我们在本文中对上面提到的问题也进行了讨论。

综合以上特点，我们认为目前中文深层语义描述主要有以下三个努力方向：

（1）根据中文语义特点，进一步规范化语义描述体系。受英文深层语义描述体系的影响，目前的研究多是对现有英文语义描述规范的继承和改造。随着中文语义理论基础和语义知识库的构建和完善，结合汉语的具体特点，探索面向中文的语义表达形式仍是一个重要问题。

（2）建立从自然语言文本到实体、概念、关系谓词之间的映射。同样的词汇和句法可以表示不同的语义；同样的语义，可以由多种词汇及句法来表达。因此，如何建立文本到语义之间的映射也是一个关键问题。

（3）对逻辑命题义描述的范围加以扩充和完

善。目前语义表示的主流研究主要集中在对句子基本的论元结构的描述上。但理论上,命题义还包括情态逻辑、时态逻辑,甚至程度等主观意义和附加意义。为便于区分,我们将这些命题语义之外的语义合称为超命题义。

2.2 中文深层语义描述体系架构

为解决上述关键问题,需要建立融合概念与逻辑的中文深层语义描述体系。我们的整体解决思路是,从计算机深度计算和语言工程的角度对文本所蕴含的语义进行分层次、细粒度的深入挖掘。向下解决词汇到概念映射的问题,借助《汉语语法信息词典》[16]和《现代汉语词典》[17]对概念义进行细化;向上从命题义向超命题义推进,突破浅层语义的局限对更广泛的逻辑义进行描述。根据以上思路我们初步建立了现代汉语谓词语义角色标注语料库规范[18]。

具体来说,三层表示机制自下而上分别为:

(1) 概念义层 描述实体概念、事件状态概念;

(2) 命题义层 描述句子的客观意义,具体对谓词事件、论元成分进行描述;

(3) 超命题义层 描述句子的主观和附加意义,具体包括情态义、时体义、程度义、否定义、情感义。

中文深层语义描述可用语义图这种图结构表示。但同时也可以拆分为若干线性表示的组合。如图 1 所示,中文深层语义描述能够将句子的深层语义表示为一个含边和节点的有向图。

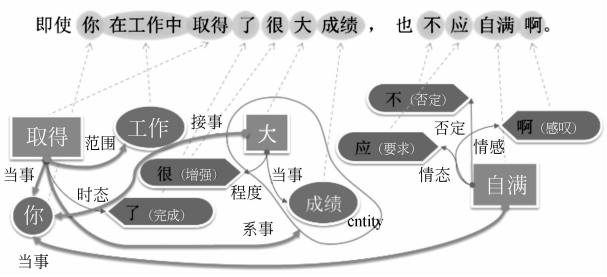

图 1 中文深层语义描述的图状表示

图 1 中,所有□或○表示句中蕴涵的概念义,其余节点则描述了一般语义角色标注中所没有的超命题义。实有向边描述了概念之间、概念与超命题义之间的语义关系;虚有向边表示语义和其在原文中的单词或语块之间的对齐关系。

中文深层语义描述还有与图结构等价的线性表示形式。例如图 1 还可以用以下纯文本的线性形式等价地表示,如例 2 所示。

例 2 S:即使你取得了很大成绩,也不应自满啊。

P1:即使[%当事你][#取得#]<义项 1><了>tense_perfect[%系事很大成绩],也不应自满啊。

P2:即使[%当事你]取得了很大成绩,也<不>logic_neg<应>mod_require[#自满#]<义项 1><啊>tone_sigh。

P3:即使[%接事你]取得了<很>dgr_high[#大#]<义项 1>[%当事成绩],也不应自满啊。

其中,S 句是原文本。P1~P3 是三个句子的部分广义命题(同时带有概念义、命题义、超命题义的描述)。每个广义命题语义表示围绕一个命题义进行描述,并附带与该命题关联的超命题义。通过构建多个广义命题义,我们就能够描述句子完整的图状结构深层语义。上例中,"< >"之后接的就是超命题义的标签类别,如 tense_perfect 表示"时态义—完成",log_neg 表示"逻辑否定"等。这种表示方法虽不如图状表示形式直观,但非常便于计算机自动处理。

此外,中文深层语义描述也可以看作若干谓词表达的组合。如例 2 也可以由下面元组的组合表示:

P1 取得(你,很大成绩),(了:tense_perfect)

P2 自满(你,自满),(不:logic_neg,应:mod_require,啊:tone_sigh)

P3 大(你,成绩),(很:dgr_high)

下面我们按照由概念义、命题义和超命题义组成的三层结构的顺序,分别介绍其具体的描述方法。

2.3 基本概念义的描述方法

在描述体系中加入概念义,是为了更加细致地展现句中蕴含的事件语义和组合语义。概念义是描述体系中语义的基本单位。基本概念义,即常见的由单词表示的语义概念。

基本概念义的描述对象是句子中的单词。区分比较易混淆的中心谓词离不开一定的语言知识库或语义词典的支持。考虑语言计算、应用的需求,我们通过《现代汉语语法信息词典》(grammatical knowledge-base of contemporary Chinese,GKB)以及第 5 版《现代汉语词典》(简称 XH5)的辅助对基本概念义进行选择和标注[19-21]。

GKB 是北京大学计算语言所为计算机实现汉语句子的自动剖析与自动生成而研制的一部电子词典,具有科学严格的收词原则,特别是语法功能和义

项相结合的原则。《现代汉语词典》是由中国社会科学院语言研究所编纂的中国第一部规范性语文词典。考虑到 XH5 的规范性和现阶段工程需求，我们将 GKB 和 XH5 的交集动词（不含其他动词、其他词类或非成词语素）的义项（15 654 个）逐条人工合并和校对，整理为新的结构化概念义数据库 BCSD。每个概念的标记格式为“拼音_词语_义项编码_义项释义_示例”。其中，义项编码是考虑标注任务需要，在数据完备整理的基础上新实现的字段。例如，“āi_挨_XH5@010201_靠近；紧接着_他家～着工厂|学生一个～一个地走进教室”，其义项编码是“XH5@010201”。在标注概念时只需标注其义项编码，例如，

例 3　他的朋友昨天从哈尔滨[＃飞＃]<％XH5@010702％>到了北京。

其中，“飞”被[＃・＃]标为谓词，后面紧跟的<％ XH5@010702 ％>是其义项编号。它表明该义项是词条“飞”在 XH5 中的第“1”次出现，该词共有“7”个义项，当前用法是其中的第“2”个义项。通过以上形式（对谓词标注义项编码）词语语义得到明确。

2.4　命题义描述方法

命题义描述的目标是将自然语言转换为某种事件框架表示，具体来说，就是句中主要动词（或形容词、状态词）跟与其共现的体词性词语（名词、体词性代词、时间词、处所词、方位词等）之间的关系所描述的基本事件语义。我们按照标注命题义三个环节（选定谓词、识别论元成分、标注论旨角色）分别进行介绍。

2.4.1　标注对象“谓词”的选择方法

句中的谓词性成分，默认都应作为标注对象。在宾州命题树库等体系中体现为动词、名词、形容词等形式[22]，本描述体系还添加了复杂的谓词性结构的描述。普通动词等形式的谓词性成分，例如 2.2 节中的例 2“即使 你 取得 了 很 大 成绩，也 不 应 自满 啊”中三个谓词“取得”“大”“自满”就都作为标注对象。为清晰起见，我们目前的标注体系规定：每个谓词标注对象占据一个文本行，因此，如果一句中有两个以上的谓词标注对象，就需要通过多次复制该句来标注其中各个谓词及其论元成分。如 2.2 节 P1～P3 所示。

语义的图状表示可以通过合并多个标注句子得到，同时完整保留了文本和标签的对应关系。研究表明，这种对应关系对后续自动语义分析标注器的训练十分重要[23]。

（1）**不作为标注对象的动词特征**　在一些特定情况下，部分谓词不适合作为语义关系标注的处理对象。这类谓词的特征是：①概念语义比较虚，在句中主要起到语法作用，句中没有体词性成分与该谓词有语义联系；②某些句法位置上的“谓词”，在句中的实际功能不是陈述一个事件，而是起指称或修饰限定的作用，因而语义上更接近体词性成分，与典型的谓词性成分的“述谓”功能有所不同。

例如“这个房间要干净些”中的“要”是典型的助动词用法，满足第一个特征；又例如，“训练大概几点开始”中的“训练”是单个动词做主语，而不是陈述现实世界中发生的具体事件，满足第二个特征，故它们均不作为标注对象。

根据以上两个特征，经过总结我们发现如下类别的动词有可能不作为标注对象：助动词、形式动词、虚化东西、泛义动词。这些在我们制订的详细规范中均分情况进行具体讨论。

（2）**复杂的谓词性结构的处理方法**　有些复杂的谓词性结构虽然表面上也满足上面的两个特征，但仍需要作为标注对象或者分情况讨论。比如述宾结构、述补结构等做主语时，该谓词性结构内部的中心谓词本身起述补作用，例如，

例 4　[＃看＃][％受事电视％]是[％施事他％]唯一的消遣。

其中，“看电视”是整句的主语，其中心谓词“看”并不是直接做主语，需要作为标注对象。

此外，“定中结构”定语位置上的谓词需要根据是否能找到有语义关联的论元成分分两种情况讨论；“并列式 VP”中两个并列项分别作为标注对象；“重叠式 VP”如“看一看”整体作为标注对象，“动结式 VP”根据是否能够分辨各自论元成分采取不同的标注方法。除此以外，还有“动趋式 VP”“离合动词”、一些比较凝固的“述补式 VP”的处理方法，限于篇幅在此不做详细介绍。

2.4.2　标注对象“论元”的选择方法

论元成分的标注有广义和狭义之分，狭义的论元成分仅限于谓词所对应事件的最简单场景中的必要参与成分；广义的论元成分则包含谓词所对应事件的真实场景中的各种可能参与成分，我们的描述体系按照广义方式标注论元，最大限度地标记句中跟谓词有直接语义联系的各种不同成分。目的是使论元成分的标注能充分反映一个谓词对应事件的各

种可能的参与成分，从而为人机问答系统提供支持。比如“吃”，真实场景中除了狭义论元考虑的“进食者”和“食品”之外，往往还涉及“工具”(如筷子)、“场所”(如食堂)、“次数”等多种有语义联系的成分，可以看作广义论元。图2为标注的核心例句集中所出现角色的频率统计图。其中，“其他”包括低频角色110种，包括“处所@受事”等组合而成的角色标签。

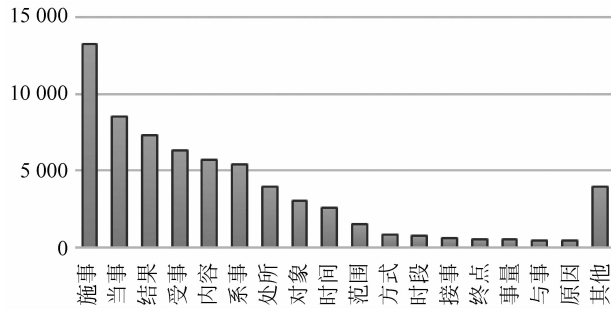

图2 各类角色在核心例句集中的频率

此外，论元成分不限于词，也可以是词组或小句。对论元成分的认定主要基于语义标准，即语义上描述了一个完整的实体、数量、时间、空间、事件(活动)等单位，构成一个“语义块”(chunk as a meaningful unit)。从句法性质上看，论元成分通常为体词性成分，比如名词性短语(NP)，还可能是介词结构(PP)、方位结构(LocP)、数量结构(NumP)等。例如，

例5 [%受事 这笔钱%]应该[%时间在一年内%][# 归还 #]。

其中，“在一年内”就是介词结构和方位结构整体充当论元。

另外也有一些跟谓词有直接语义联系的成分是谓词性结构，大多是动词性短语(VP)，应标记为[%VP角色名称 %]。也有的形容词性成分(AP)充当论元，则标记为[%AP角色名称 %]。例如，

例6 [%VP 当事 这位 同志 办事 %][# 认真 #]。

例7 [%当事 她 %][# 感到 #][%AP 内容 幸福 和 骄傲 %]。

确定论元成分时还会面临若干特殊问题，例如，不连续成分、复指成分、省略成分的结构、身份待定、论元成分的嵌套等。在语义标注规范中，我们为这些特殊情况定义了特殊标记，尽可能把句中的论元成分都纳入到句中谓词的论旨角色体系中。如对复指成分充当论元的情况：

例8 a：[%受事 这些人 %][%施事 我们 %][# 聘 #][%& 受事 他们 %] 来当顾问。

b：[%当事 这些人 %]我们 聘[%& 当事 他们 %] 来[# 当 #][%系事 顾问 %]。

句中，普通名词性成分按照一般的语义角色进行标注，而起复指作用的代词成分则用[%& %]加以标记。

其他标记详见描述规范[18]。

2.4.3 论旨角色标注的处理策略

在参考相关文献[24]对谓词论旨角色的分类基础上，我们提出了用于标注汉语句子谓词的论旨角色体系，共包含有28种围绕谓词的论元角色。

按照该成分跟谓词语义关系的紧密程度不同，可以分为核心角色和外围角色两大类。如图3所示。

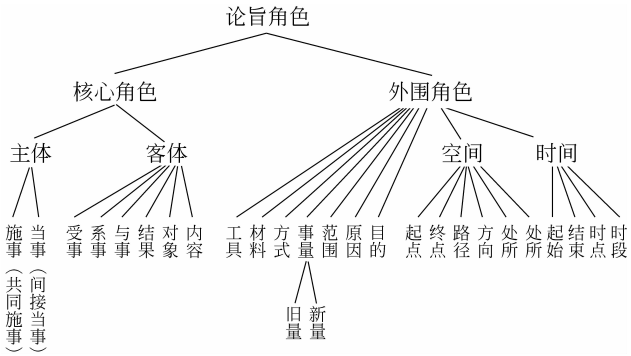

图3 中文深层语义描述论旨角色体系

一般而言，论旨角色标注应遵循“论旨角色唯一性原则”，即一个谓词在句中不应有两个论旨角色完全相同的论元成分。但实际情况中有大量难以给出明确角色判定的情况。例如：

例9 a：王老师在考小明。

b：小明在考语文。

例9(b)中“小明”跟“考”是什么关系？“语文”跟“考”又是什么关系？就不太容易确定。例9(b)中的“小明”和例9(a)中的“小明”相对于“考试”事件来说，扮演的应该说是相同的语义角色，但如果把例9(b)中的“小明”的语义角色分析为“受事”不是很典型，因为例9(b)中的“小明”即使作为“受事”，也没有明显的“被动性”。另外，如果“小明”是主动参加考试，那例9(b)中的“小明”是施事，还是受事呢？

宾州树库和抽象语义表示标注原则规定：在遇到角色重合或模糊等情况时，遵循核心角色优先原则，中文动词词汇语义网遵循框架[25]为本构成为用。为兼顾语义分析的精度与语言工程工作量之间

的合理平衡，在对大规模真实语料做论旨角色标注的具体实践中，我们主要遵循如下处理策略。

(1) 典型范畴策略

如果一个句子中的某个论元成分的角色归属不够清晰，比如既像“受事”，又像“对象”，则应该遵循典型范畴原则，即其论旨角色的判定应尽量符合某种论旨角色的典型特征。在 A、B 两个角色中做选择时，应着重考虑当前论元成分的属性特征是接近 A、B 二者中哪一个角色的典型特征。最终论旨角色的归属应该是符合典型特征多的那一个角色。

(2) 角色半开放策略

在规范给出的初始论旨角色标签系统之外，允许标注人员根据自己的认识，添加新的语义角色标签(即“用户自定义标签”)，对现有的标签做出更为详细的说明。在图 3 给出的论旨角色基础上，标注人员可以对某一个论旨角色进行细分，在基础论旨角色标签之后，用 * * 标记标注人员自定义的论旨角色，如“[%施事 * … * %]”。

在 * * 中标记的论旨角色可以给出一个初始的建议集，例如，“批评者、支持者、拥有物、……”。该论元成分作为这个角色范畴的一个成员，并不具有典型性。从某种程度上说，也可以认为是暂且搁置疑难问题，留待将来对这类论元成分的角色归属做进一步的研究。

(3) 标注粒度弹性策略

图 3 中的叶子节点(如“施事、当事、起点、终点”等)是进行谓词论旨角色标注时应优先选择的标签。例如，如果能判定一个论元是“施事”，就不应该标为“主体”。但是，在真实语料标注时，如果一个论元成分确实无法判定为“施事”或“当事”，可以回退到二者的上层节点，判定为“主体”。回退的论旨角色标注只允许到“主体、客体、时间、处所”这个层次。

2.5　超命题义描述方法

尽管目前语义分析研究还主要集中在对句子基本的客观语义即命题义的描述上，命题义理论上包含的情态逻辑、时态逻辑，甚至程度等主观意义和附加意义极少被实际标注。为了对逻辑命题义描述的范围加以完善，除沿用传统语义角色(包括施事、受事、当事、系事等)外，我们将情态义、时体义、否定义、程度义、情感义纳入“超命题义”，融入到深层句义的描述对象中，并分别对它们进行分析。

2.5.1　情态义

在模态算子后面一律加上 mod_category 作标志，并用花括号标志其辖域，这样，不仅标明了其所支配的动词性成分的范围，而且还标明了花括号中的语言表达在情态上表示的是一种非现实的断言(irrealis assertion)。例如，

例 10　农民们非常＜乐意＞mod_intention 地{帮助了我们}。

其中，mod_intention 是情态义的一个子类“意愿”。

2.5.2　时体义

“时”用来称呼具体的时态；“体”是用来描写动作行为进行状况。时体算子包括“将、刚、刚刚、已经、曾经、又、再、正、在、正在”等时间副词、“着、了、过”等时态助词、“了、呢、着呢、来着、来的”等语气词。其中，“将、即将、再”等表示将来时(future，简写为 tense_fut)；“刚、刚刚、已经、曾经、又” 等表示过去时(past，简写为 tense_past)，；其他还包括：进行体(progressive aspect，简写为 tense_prog)、完成体(perfect aspect，简写为 tense_perf)、现在完成体(present perfect，简写为 tens_pres_perf)、过去完成体(past perfect，简写为 tense_past_perf)。例如，

例 11　快乐 ＜在＞tense_prog 等待 我们。

其中“＜在＞tense_prog”表示阶段一直在进行。

2.5.3　否定义

否定算子(negative operator，neg)主要是副词“不、不必、没、没有、未、未曾”等。标注语料中，在否定算子后面一律加上 neg 作标志，并用花括号标志其辖域。例如，

例 12　就我自己的愿望来说，我连一天也＜不＞neg{想 呆}。

2.5.4　程度义

程度算子(degree operator，dgr)主要是指副词“很、非常、特别、蛮、过分、最、不大、稍微、稍许、有点儿”等，对于修饰的主要谓词限制程度。标注语料中，在程度算子后面一律加上 dgr_high 或 dgr_low 作标志，并用花括号标志其辖域。例如，

例 13　理论与实际的结合是＜非常＞dgr_high{紧密}的。

2.5.5　情感义

情感义是说话人对句子韵律特征进行加工而表达的体现自己交际意图的主观念。目前学者们在语气是否属于模态范畴这个问题上还存在争议：台湾

中研院的谓词语义角色体系中有评估语气、感叹词、句尾语气、选择语气等角色，贺阳[26]把“Modality”称之为“语气”，张喜洪认为意念系统中有三个子系统：情态、语气和口气。本文提出的描述体系中，情感义仅包含狭义的功能语气，按表达的语气分为陈述语气、疑问语气、祈使语气、感叹语气四大类，分别标记为：intonation_statement、intonation_question、intonation_imperative、intonation_exclaimation。例如，

例14 班长的学习是全班最好＜的＞intonation_statement。

并对表现各种功能语气的词语以“词语”“词性”“ID”“释义”“用法”“典型例句”和“语气类别”来描述它们。

超命题义的标注虽可囊括大部分汉语语义现象，但在理论上，仍有一些问题存在争议或者模糊难以判别，如虚拟语气(subjunctive mood)用于表示一种假想的情况或主观愿望。是否单列出来？“不必”和“不用”，“用”是许可还是要求？为保证现有体系的准确性，我们暂时不做标注，留待以后进一步讨论。超命题义的完整分类标记集见附录表格1。

3 基于中文深层语义描述的语言资源构建

语料库对于自然语言处理研究的巨大价值已经得到学者们认可。语义标注的语料库构建目前主要以人工标注为主，如宾州树库、框架语义网、抽象语义表示等。传统人工标注的优点是在标注量小的情况下准确性高，但标注的一致性、进度、质量受到标注者相关因素影响较大，难以高效应对大规模语料库的标注需求。

为了获得高质量的中文深层语义语料库，同时避免传统人工标注的低效问题，我们采用社会标注(social annotation)中基于群体智慧(collective intelligence)语料标注方法对语料资源进行标引、组织和标注[27]。基于群体智慧的标注与其他社会化标注方法(如众包标注等)不同的是：该方法不只将标注任务分工，更强调对标注参与者的智慧的运用和发掘，以及对其结果的有效归纳、加总，以形成最终的集体性成果，因此更加适合对准确性和专业性要求较高的语义语料库标注需求。图4描画了我们所使用的基于群体智慧的标注模型。

该模型的输入项有三个：标注者、待标注语料和标注规范。模型的处理部分包含：能力评测、语料筛选、任务组织等，输出为标注语料库。其中核心处理部分是：语料筛选、协作标注和决策加总。

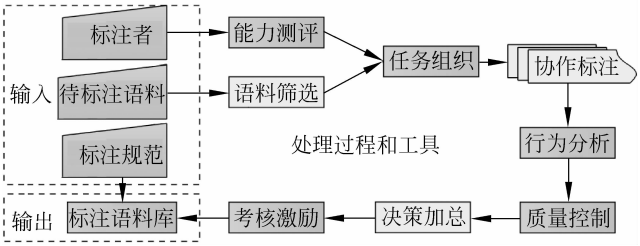

图4 基于群体智慧的标注模型

3.1 语料筛选

就语料库标注来说，一个重要的指标是语料的平衡。平衡语料能够更好地覆盖各种语言现象，并减轻数据稀疏问题。语料平衡需要考虑多个因素(如语义角色标注中，谓词、句式、意义组合模式等都可以作为参考)，因此标注者的专业知识和经验对于平衡语料十分重要。在基于群体智慧的语料标注过程中，我们通过专家级标注者提供语料特征分析、语料检索、语料统计、词典对比、语料抽取等工具，并由专家级用户挑选代表性语料进入候选标注语料。

最终用于标注中文深层语义的原始语料由核心例句集(ICL平衡例句集)12万字以上、常用动词例句集31万字以上、《人民日报》语料113万字以上、微博和网络语料11万字以上，以及多领域文本语料800万字左右构成(包括微信公众号文章、小说散文、社会科学、自然科学、中学课本、科普类文章)。所有语料合计超过1 000万字。其中核心例句集是经过精加工的高质量标注数据集，其余部分是以机器自动标注为主，人工为辅完成的，在后续阶段将会继续完善。

3.2 语料标注

收集到原始语料后，语料标注的核心就在于协作标注模块。为了减轻用户决策受到互相的影响，我们基于隔离标注法，即多个用户标注同样的语料，彼此之间无法看到对方的标注结果，设计了专用标注平台，并收集有差异的标注进行典型差错分析，进而改进标注规范和标注工具。

此外，从语言工程角度讲，把句子语义的分析分解为多个层面分阶段处理，更有利于把握分析质量、内部一致性，控制工程进度，因此本课题规定标注者采用分层标注、层层渲染的方式进行句子语义标注，即每一层标注都针对句子三层语义架构中的一个层

面的(局部)问题，最终汇聚成对句子语义的完整描述。

3.3 语料聚合

为了获得最终的一致的集体性标注成果，我们还需要进行语料聚合——设计有效的决策机制，对群体的个人智慧进行有效加总。决策加总机制实现为一个信息聚合模块，包含三个单元：生成方案单元、优化方案单元以及评估方案单元。同时在生成、优化以及评估方案中，我们选择“外扩”的方式弥补决策偏差。相对自组织、加权平均，外扩实施起来困难最大，但最为有效[28]：即在收集和评估决策方案的时候，去寻找外界的帮助，扩大参与决策的个体数量。

在我们的项目实践中，基于群体智慧的标注方法在面对一定规模的语料标注任务时，比传统的手工标注单人标注速度提升至 1.5 倍，同等数量语料标注速度提升至 7 倍，标注质量也有所提升。其原因在于：在创新性很强的自然语言处理项目实施过程中，其探索的特点非常明显。就我们的标注任务来说，早期标注规范是随着语料标注的进展逐步提炼、修改、完善，这个过程需要有效的群体协作、信息发掘、智慧归总，才能最终形成高质量的大规模标注成果。这种情况下，基于群体智慧的标注方法相比传统方法有明显的优势。

4 结论

本文针对中文深层语义描述及其资源构建进行研究，研究工作及所取得的成果可以概括为以下三个主要方面。

(1) 本文结合汉语本身特点，提出以基本命题义为出发点，向下融入概念义、向上融入超命题义的中文深层语义描述体系规范。

(2) 基于相关理论和大量文本实例，对基本命题义的标注对象(谓词性成分和论元成分)和论旨角色重新定义，并对汉语中的特殊现象进行单独分析，并定制多种标注策略。

(3) 在中文深层语义描述体系的基础上，建立了完善的语料标注模型，通过自建平台实现了快速、高质量的大规模语料人工标注。

本文的特色在于，从中文“意合”的语言特点出发，同时将语言学理论与计算机工程相结合，提出结合概念和逻辑的、涵盖命题语义和超命题语义的中文深层语义描述体系，可以真正实现分层次、细粒度挖掘汉语文本中的语义信息。此外，在设计时就将自动分析自动推理作为考虑的首要因素，在概念义、命题义、超命题义的表示上既保证了语义图结构的完整性，同时保护了文本和标注结果的关联关系，能够直接为自然语言处理和理解的多个研究领域(如词义消歧、机器翻译、信息抽取和大规模语料库加工、归纳和推理)提供较为全面、深入的语义知识，为自动语义分析提供更为充分的支持。

本文虽然在中文深层语义描述体系和资源的建设方面取得了一定成果，但是离实用化的目标还有很长的路要走，如下问题均可进一步展开：如何提高自动中文深层语义分析器的性能；如何对概念义涉及的短语、单词嵌入的进行学习；如何解决隐式语义角色标注的问题。此外，非规范文本的语义标注，如微博等社交媒体网站产生大量的口语化、弱规范甚至不规范的短文本，在标注时的速度和质量都相对较低，该如何解决？

参考文献

[1] Kingsbury P, Palmer M. From TreeBank to PropBank [C]//Proceedings of the LREC, 2002: 1989-1993.

[2] Fillmore C J, Johnson C R, Petruck M R L. Background to FrameNet [J]. International Journal of Lexicography, 2003, 16(3): 235-250.

[3] Liang P, Jordan M I, Klein D. Learning dependency-based compositional semantics [J]. Computational Linguistics, 2013, 39(2): 389-446.

[4] Oepen S, Kuhlmann M, Miyao Y, et al. SemEval 2014 Task 8: Broad-coverage semantic dependency parsing [C]//Proceedings of International Workshop on Semantic Evaluation, 2015: 63-72.

[5] Oepen S, Kuhlmann M, Miyao Y, et al. Semeval 2015 task 18: Broad-coverage semantic dependency parsing [C]//Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), 2015: 915-926.

[6] Che W, Zhang M, Shao Y, et al. SemEval-2016 Task 9: Chinese semantic dependency parsing [C]//Proceedings of Joint Conference on Lexical and Computational Semantics, 2012: 378-384.

[7] 刘挺，车万翔，李正华. 语言技术平台 [J]. 中文信息学报，2011, 25(6): 53-63.

[8] Che W, Zhang M, Shao Y, et al. SemEval-2012 Task 5: Chinese semantic dependency parsing [C]//Proceedings

of Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics, 2012:378-384.

[9] Banarescu L, Bonial C, Cai S, et al. Abstract meaning representation for sembanking[C]//Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, 2013: 178-186.

[10] Hovy E, Marcus M, Palmer M, et al. OntoNotes: the 90% solution[C]//Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics, 2006: 57-60.

[11] Li B, Wen Y, Weiguang Q U, et al. Annotating *the Little Prince* with Chinese AMRs[C]//Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016), 2016: 7-15.

[12] Zelle J M, Mooney R J. Learning to parse database queries using inductive logic programming[C]//Proceedings of the National Conference on Artificial Intelligence, 1996: 1050-1055.

[13] 顾阳. 论元结构理论介绍[J]. 当代语言学, 1994 (1): 1-11.

[14] 吴平. 试论事件语义学的研究方法[J]. 外语与外语教学, 2007 (4): 8-12.

[15] 陆俭明. 词语句法, 语义的多功能性: 对“构式语法”理论的解释[J]. 外国语, 2004, 2(2): 15-20.

[16] 俞士汶, 朱学锋, 王惠, 等. 现代汉语语法信息词典规格说明书[J]. 中文信息学报, 1996, 10(2): 1-22.

[17] 现代汉语词典[M]. 北京: 商务印书馆, 2002.

[18] 詹卫东, 穗志方, 常宝宝, 等. 现代汉语谓词语义角色标注语料库规范[EB/OL]. [2018-07-12]. http://www.klcl.pku.edu.cn/xwdt/231644.htm.

[19] 邱立坤, 赵慧, 俞士汶, 等. 《现汉》与《语法信息词典》词类对应分析[J]. 中文信息学报, 2017, 31(5): 1-7, 20.

[20] 傅爱平. 汉语信息处理中单字的构词方式与合成词的识别和理解[J]. 语言文字应用, 2003 (4): 25-33.

[21] 詹卫东. 面向自然语言处理的现代汉语词组本位语法体系[J]. 语言文字应用, 1997, 4: 101-106.

[22] Xue Nianwen, Martha Palmer. Adding semantic roles to the Chinese Treebank[J]. Natural Language Engineering, 2009, 15(1): 143-172.

[23] 袁毓林. 谓词隐含及其句法后果——“的”字结构的称代规则和“的”的语法、语义功能[J]. 中国语文, 1995 (4): 241-255.

[24] Wang C, Xue N. Getting the most out of AMR parsing[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 1257-1268.

[25] 刘美君, 万明瑜. 中文动词及分类研究: 中文动词词汇语义网的构建及应用[J]. 辞书研究, 2019, 2.

[26] 贺阳. 试论汉语书面语的语气系统[J]. 中国人民大学学报, 1992(5): 59-66.

[27] 张喜洪. 现代汉语情态范畴初论[D]. 成都: 四川师范大学硕士学位论文, 2008.

[28] 柯永红, 俞士汶, 穗志方, 等. 基于群体智慧的语料标注方法研究[J]. 中文信息学报, 2017, 31(4): 108-113.

[29] Elizabeth M Daly. Harnessing wisdom of the crowds dynamics for time-dependent reputation and ranking [C]//Proceedings of International Conference on Advances in Social Network Analysis and Mining, IEEE, 2009: 267-272.

夏乔林(1992—), 博士研究生, 主要研究领域为自然语言处理。
E-mail: xql@pku.edu.cn

常宝宝(1971—), 博士, 副教授, 主要研究领域为自然语言处理、计算语言学。
E-mail: chbb@pku.edu.cn

穗志方(1970—), 通信作者, 博士, 教授, 主要研究领域为计算语言学、知识工程。
E-mail: szf@pku.edu.cn

附录

表格 1　超命题义分类标记集

类别			典型词语	标注标记
情态	可能	必然	一定、必然	mod_certainty
		或然	可能、也许	mod_possibility
	能愿	意愿	肯、不好不	mod_intention
		能力	能、会	mod_ability
	履义	许可	可以、能	mod_permission
		要求	应、必须	mod_requirement
	评注	领悟	原来、怪不得	mod_comment_1
		料定	果然、果真	mod_comment_2
		庆幸	幸亏、幸好	mod_comment_3
		意外	居然、竟然	mod_comment_4
		情绪	老是、本来	mod_comment_5
	建议		最好	mod_advice
	评判		值得	mod_judgement
	反诘		何必、何不	mod_rhetorical
	捩转		反而、反倒	mod_torsion
	强调		的确、就	mod_emphasis
时体	将来时		将、即将	tense_fut
	过去时		刚、刚刚	tense_past
	进行体		着、正	tense_prog
	完成体		了、过	tense_perf
否定			不、没有	neg
程度	增强程度		很、非常	dgr_high
	削弱程度		稍微、稍许	dgr_low
语气	陈述语气		的、了	intonation_indicative
	疑问语气		吗、呢	intonation_interrogative
	祈使语气		吧、嘛	intonation_imperative
	感叹语气		啊、呀	intonation_exclaimation