汉语句法语义链接知识库开发

周 强

智能技术与系统国家重点实验室
清华大学计算机系, 北京 100084
zq-lxd@mail.tsinghua.edu.cn

摘要

本文以汉语中描述存在状态和拥有关系及其变化转移的一组典型动词作为研究对象，提出了一套构建大规模汉语句法语义链接库的解决方案：首先融合现有句法描述资源，开发词汇关联知识库；然后设计情境语义描述体系，开发针对汉语拥有和存在类动词的情境语义知识库；最后使用情境语义知识库中的语义描述信息对词汇关联库的各个相关词汇对进行语义信息标注，形成句法语义链接知识库。该方案在单义存在类动词关联对的参量锚定处理中已取得初步实验效果。

1 引言

句法语义链接（Syntax-Semantics Linking），是自然语言理解中的一个非常重要的研究课题。它需要解决句子表层的主、状、宾等句法功能成分与深层的逻辑主语和逻辑宾语之间的链接关系，涉及到以下技术难点：1）相同句法结构的不同语义解释，其中可能涉及词汇语义及搭配因素；2）相同语义结构的不同句法实现等。其中最重要的是对句子中以动词为中心的基本事件内容的分析及其相关事件描述单元的锚定处理。

对此，语言学家的解决方案是对不同语言的句法语义链接特点进行深入分析，提炼出一些通用的句法语义链接原则和处理规则(Vanlin and Lapolla, 1997)；计算语言学家的解决方案是通过对大规模真实文本的句法语义标注建立两者之间的内在联系，典型的例子包括英语的 PropBank (Kingsbury and Palmer, 2002)和 FrameNet (Fillmore and al. 2001)项目等，并以此为基础训练不同的统计分析模型，初步实现了对英语文本的自动语义角色标注(Gildea and Jurafsky, 2002)。

我们的解决方案则是在特定的词汇关联对上，同时描述它在真实文本中可能形成的句法组合关系和语义角色关系，从而在词汇层面上直接建立起词汇对的句法语义链接关系。以此为基础构建大规模的汉语句法语义链接知识库，可以为汉语真实文本的句法语义自动提供有力的支持：一方面，将其中大量可靠的词汇关联对应用于句法结构排歧，可以不断提高分析器对真实文本语料的处理性能，同时通过其内置的句法语义链接关系直接得到相应的语义关系描述；其次，链接知识库中丰富的词汇组合实例，又可以为语义抽象和概念描述提供重要的语言应用素材，通过建立它们与不同语义资源的内在联系，可以不断扩充现有链接库的知识容量，缓解词汇层面的数据稀疏问题对真实文本分析应用的影响。

从这个设计思路出发，本文以汉语中描述拥有关系和存在状态及其变化转移的一组典型动词，如：有、存在、送、创造等作为研究对象，提出了一套构建大规模汉语句法语义链接库的解决方案：首先融合现有句法描述资源，开发词汇关联知识库；然后设计情境语义描述体系，开发针对汉语拥有和存在类动词的情境语义知识库；最后使用情境语义知识库中的语义描述信息对词汇关联库的各个相关词汇对进行语义信息标注，形成句法语义链接知识库。下面几节将对有关内容进行详细介绍。

2 词汇关联知识库

我们目前的词汇关联知识库主要描述汉语中任意两个实义词在真实文本句子中可能形成的各种句法组合实例，内容包括：定中、状中、述宾、述补、主谓、并列、连谓和介宾等 8 种句法关系。主要数据来源是以下几个语言资源库：

1）语义关联网：从现代汉语辞海(张卫国等,1994)中自动提取出的各种实词搭配对，覆盖上面定义的前 7 种句法关系，并对其中的所有词

语进行了同义词词林的语义代码自动标注(苑春法等，1997)。其特点是所有词汇对由语言学家根据自己的语感进行总结提炼，因此可靠性较高，同时给出了同一词汇对的各种可能的句法关系组合，为自动分析它们在句子中的可能变位情况提供了重要支持信息；

2）清华句法树库：在 100 万词的覆盖文学、新闻、学术和应用等体裁的汉语平衡语料库的真实文本句子上自动标注了完整的句法结构树信息（周强,2004）。利用其中的语法关系标记，可以自动提取出上面定义的 8 种关系的词汇对描述实例，形成完整的真实文本词汇关联对分布数据；

3）语义骨架标注语料库：在人民日报 1998 上半年的 2 个月的语料文本上人工标注了句子的语义骨架信息（因事 S-动词-果事 O），从中可以自动提取出两大类词汇关联关系：“S-V”和“V-O”。由于其中包含了初步的语义信息，可以在句法语义链接中发挥作用；

4）北大语法信息词典(俞士汶等，1998)：从中可以提取了以下两类词汇关联对：述补关联对和量名关联对。其特点是由语言学家进行了全枚举分析，因此数据比较完整全面，准确性较高。

将以上不同来源的词汇关联对数据进行汇总，共得到约 96 万个词汇关联对描述，它们形成了进行句法语义链接处理的基础数据。下面列出了从中提取出的动词‘有’相关的几个关联对实例：

- 有　火车　PO TCT　V N　*　1
- 有　火点　ZW SKT　V N　*　1

其描述格式为：<中心词语> <关联词语> <关系标记> <来源信息> <中心词类标记> <关联词类标记> <分隔信息> <出现频度>[1]。

通过建立其中的各个关联对与不同层次的标注语料库中的相应句子的联系索引，可以形成配套的文本句子库。目前主要包括以下内容：句法树库中标注有完整句法树的切分和词性标注句子，语义骨架库中标注有语义块和中心词的切分和词性标注句子以及人民日报库中标注有切分和词性标记的句子。以后需要时还可以不断补充其他切分和词性标注语料库，并通过启动自动分析器支持下的词汇关联知识自动获取工具从海量的真实文本语料库中不断提取比较可靠的词汇关联对和句法标注句子补充入相应的知识库中，从而形成足够规模的用于进一步的句法语义链接处理的词汇句法关联数据。

3　情境语义描述体系

情境语义描绘体系的设计目标是对汉语中的 V 类事件概念和 N 类实体概念进行概要描述，形成句法语义链接处理的基础词汇语义描述数据。

在事件概念描述方面，主要采用了情境表达式和参量锚定机制相结合的处理策略。其中，情境表达式采用二阶谓词逻辑描述形式，通过引入谓词参量，提高了对不同事件内容描述的灵活性和有效性。参量锚定机制则是将情境表达式中的不同参量与真实文本句子中的主要信息描述单元，包括 V 块、N 块等联系起来的重要手段，据此可以方便地建立起抽象的情境表达式与具体的句子描述实例之间的内在联系，形成针对不同句子描述的事件内容的完整解释。下面给出两者的基本定义方法：

情境表达式由基本事件描述式和运算符组成，其中：基本事件描述式采用了常用的谓词-论元结构来描述基本事件内容，具体定义式为：<基本事件描述式> ::= <谓词> (<论元 1>, <论元 2>, … <论元 n>)。通过定义不同的谓词和参量可以对不同的动态动作行为和静态状态关系进行描述。而引入运算符的作用则是将不同基本事件描述按照其逻辑关系组织起来，形成复杂情境描述的有机整体。目前主要定义了以下几个运算符：1）NOT：逻辑非；2）&：逻辑与；3）CAUSE：逻辑蕴涵。

在实体概念描述方面，则提出了以人为本的基于物质世界、精神世界、符号世界和人际社会等四个世界划分方法的平面网状结构的基本语义类组织策略。其中：

物质世界：由物质实体构成，在一定的时间里占据一定的空间。其发展变化和相互作用一般都伴随着能量转换过程。物质实体都有自然与人造的区分，它们构成了人的外部环境的主体部分。

精神世界：由心理状态、心理感受与心理活动及其产物构成，是影响人的行为、行为方式、行为结果的内在因素。

符号世界：由信息的构成要素：符号体系、符号载体和物质实现三部分组成，是人对外部世界认知的不可或缺的重要部分。

人际社会：由人、人群、机构法人等组成，形成人类的基本生存环境。人类社会中存在的各种实体、关系、状态、事件、现象、活动、行为等与物质世界、精神世界、符号世界中的对象相比

[1] 其中使用的若干符号简要说明如下：关系标记：PO—述宾，ZW—主谓；来源信息：TCT—清华句法树库，SKT—语义骨架库；词类标记：V—动词，N—名词；分隔信息：*--空分隔成分。

具有不同的特性。

以此为基础，我们初步总结了反映四个世界基本特征的人工物、自然物、精神、信息、组织、人和事件等基本语义类，再加上属性、数量、时间和空间等概念，形成了反映人类基本认知过程的基本概念描述体系。

这些上层的基本 V 类事件概念和 N 类实体概念与上节给出的下层的词汇关联对和真实文本句子相配合，可以为针对特定语义描述问题的词汇语义计算研究提供巨大的灵活性和广阔的内容提升空间。

4 情境语义知识库

我们目前的描述重点是汉语中表示拥有关系和存在状态及其变化转移的相关动词和典型句式，希望利用上面提出的情境语义描述体系，对其中的每个动词和句式给出合适的情境表达式描述，形成相应的情境语义知识库。

为实现这个目标，我们提出了以下处理策略：首先通过人工思辨，归纳总结针对拥有和存在类描述的基本谓词和导致它们产生转移变化的扩展谓词，深入它们之间的内在联系，对这些谓词给出完整的情境表达式描述，形成情境语义 Schema；然后，充分利用现有语义描述资源，确定 Schema 中每个谓词可能包含的动词词表，通过将 Schema 中不同谓词的情境语义描述信息赋值给其中的每个动词，形成针对各个动词的情境语义知识描述；最后，人工分析形成各个典型句式的情境语义描述。下面以存在类动词为例对此进行简要说明：

在存在状态方面，我们首先区分出以下两个本原谓词：1）实体存在状态：exist(x,L)，其中的经历主体 x 主要包括具体物、精神、信息和信息载体等；2）生命期存在状态：exist-t(x,L)。它强调了存在状态的时间轴描述特点，其中的经历主体 x 主要包括人、组织和事件等。从这两个本原谓词出发，可以分别引申出一组基本谓词和相关的动作行为谓词，形成两个紧密相关的存在状态描述的情境语义 Schema。

图 1 显示了其中的实体存在状态的情境语义 Schema。从本原谓词 exist(x,L)中引申出以下两个基本谓词：1）出现状态 appear(x,L)，强调了从无到有的状态变化后的存在状态，即：appear(x,L) ::= do(x,~) CAUSE exist(x,L)；2）消失状态 disappear(x,L)，强调了从有到无状态变化后的存在状态，即：disappear(x,L)::= do(x,~) CAUSE NOT exist(x,L)。

在此基础上，可以进一步总结出导致以上 3 个状态发生变化的相关动作谓词。其中，针对 exist 状态的主要有：创生物质主体的创建(Create)和制造(Produce)谓词等，创生信息主体的写作(Write)和打印(Print)谓词等，使相关主体消失的删除(Remove)和摧毁(Destroy)等谓词，以及比较特殊的伪造(Forge)谓词等。针对 appear 状态的主要有：使物质主体出现的揭露(Reveal)和剥去(Strip)等；使信息主体出现的发表(Announce)和提出(Propose)等。针对 disappear 状态的主要有：使物质主体消失的藏匿(Hide)和覆盖(Cover)等；使信息主体消失的隐瞒(Conceal)、不说(KeepSilence)等。而汉语中的存在句式：S V N，则与这 3 个状态建立了紧密的内在联系，成为描述实体存在状态出现和消失的典型句子语境。

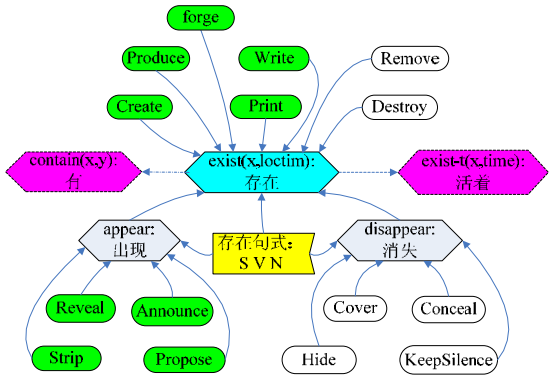

图 1 实体存在状态描述 Schema

而其中的本原谓词 exist 又可以与生命期存在和拥有关系 Schema 中的本原谓词 contain 和 exist-t 建立起紧密的内在联系，从而形成一个针对汉语拥有关系和存在状态的完整情境语义描述 Schema 网络。

然后，我们利用现有语义资源描述词典，包括知网和词林等，使用 Schema 信息对其中的相关语义类进行情境定义和描述。这样，以这些语义类标记为中介，就可以建立起各个语义类中的动词义项与相应的情境表达式之间的内在联系。例如：我们对知网中的义原类“create|创造”给出以下的情境描述：DO(x,P(x,y)) CAUSE exist(y,L) [P=Create]，就可以得到以下相关动词的情境描述信息：

- 创造 DO(x,P(x,y))_CAUSE_exist(y,L) [P=Create]
- 生成 DO(x,P(x,y))_CAUSE_exist(y,L) [P=Create]

5 句法语义链接知识库开发

句法语义链接知识库的开发目标，是利用情境语义库对词汇关联库中的关联对进行情境表达式和参量锚定信息标注，在词汇层面上建立起句法关系与谓词-论元结构之间的内在联系。

这里的主要处理难点是：1）关键动词的情境语义确定，需要顺序解决以下问题：a）关键动词是否描述了我们目前关心的存在/拥有类情境？b）具体应该选择哪个情境表达式进行义项标注？2）关联词语的参量锚定，需要准确分析关联词语与关键动词之间的不同语义关系，使用情境表达式中的合适参量标注其核心语义联系，使用其他标记标注其他可能的句法语义联系。

我们的基本方法是：以词汇关联对描述信息为中心，充分利用其中的词汇描述和句法分布信息，并参考相应的标注语料库句子，开发半自动的语义信息标注工具，从单义词对到多义词对，从易到难地逐步扩大句法语义链接知识库的标注规模。在链接知识库达到一定规模时，适时启动词汇语义内容计算过程，不断调整现有的情境语义 Schema 描述，使之达到更好的描述效果。最后，利用各个关联对相关联的文本句子索引，将词汇关联对层面的语义标注信息直接映射到真实文本句子，完成大规模语料库的句法语义信息标注。具体处理步骤如下：

首先以情境语义库的动词词条为关键字检索词汇关联库，发现所有包含这些动词的词汇关联对，从中提取与情境描述密切相关的句法关系描述，包括：定中、述宾、主谓和以介词短语作状语的状中结构，按照其中关键动词的不同义项分类组织成以下不同的初始链接知识库：1）绝对单义：在现有描述体系下只有一个义项，并且它是存在或拥有类义项；2）存在/拥有类单义：在现有描述体系下存在多个义项，但只有一个我们关心的存在/拥有类义项；3）存在/拥有类多义：在现有描述体系下存在多个义项，同时在我们关心的存在/拥有类中也存在多个义项。

然后，顺序进行以下阶段的处理：

1）绝对单义动词链接库的参量锚定处理：重点解决其中的参量锚定问题，通过自动参量锚定和人工校对，构建一定规模的完整句法语义链接库。这里的基本处理原则是：通过不同锚定参量反映关联词语与关键动词之间形成的可能语义联系的强弱程度。其中：

a）最强的语义联系是关联词语充当关键动词反映的情境描述的核心语义角色，判定方法是将它们代入关键动词对应的情境表达式的相应参量位置时可以形成合理的情境语义内容解释。此时，就可以用相应的参量标记：x，y 或 L 等对此进行标注，实现参量锚定；

b）较弱的语义联系是关联词语充当关键动词反映的情境描述的外围语义角色，表示事件内容实现的方式、工具、原因、材料等。简单判定方法是将它们代入不同的介词格标鉴定式中可以给出合理的语义解释：p <关联词语> <关键动词>。此时，可以用 O 标记进行标注。

c）更弱的语义联系是两者之间形成一种修饰和被修饰的关系。此时，可以用 Q 标记进行标注。

2）基于单义动词链接信息的词汇语义本体计算：对实现了句法语义链接分析的词汇关联对进行各个典型情境类的数据汇总和基本语义类抽象分析，从中提取情境语义区分度较高的典型词汇和语义类信息组成特征描述向量，构建针对不同存在/拥有类情境的词汇语义本体（Lexical Semantic Ontology，LSO）。在此基础上，可以启动基于 LSO 的自动聚类分析，据此实现对初始情境语义 Schema 的动态调整和内容完善处理。

3）基于词汇语义本体的多义动词义项确定和参量锚定处理：充分利用自动计算得到的 LSO 中的情境语义区分度较高的典型词汇和语义类信息，与各个多义动词关联对进行相似度匹配计算，从中选择置信度较高的匹配结果，利用 LSO 中提供的情境语义描述和参量锚定信息，完成针对这个关联对的相关语义信息标注。在此基础上再进行必要的人工校对，可以大大提高对多义动词关联对进行语义信息标注的处理效率。

4）真实文本句子的句法语义信息标注：将关联对中的语义标注信息直接映射到其对应的语料库标注句子上，在句子的各个典型句式中发现可能的语义信息冲突现象，利用不同语义组合优先度进行选择排歧，实现真实文本句子的句法语义标注，构建完整的句法语义标注语料库。

以上处理步骤顺利实现的关键是开发一组有效的词汇语义计算工具，包括单义动词关联对的自动参量锚定工具、基于单义动词链接信息的 LSO 计算工具、情境语义自动聚类和 Schema 重组分析工具、多义动词关联对义项排歧和参量锚定工具、真实文本句子语义信息标注工具等。这些工具将集成在一个功能强大、方便易用的人机交互的句法语义链接知识库开发平台中，通过为标注人员提供各种全方位的词汇语义信息计算支持，便于他们对某个关联对及其关联句子的语义信息标注和校对形成准确判断。

6 初步实验结果

6.1 基础资源

我们首先以知网 2000 版数据为基础生成了以下情境语义知识库：动词词条 3380 个，相应义项 3678 个，涉及拥有关系(H)、实体存在(E)和生命期存在(L)的义项总数分别为 1400、1384 和 894 个，发生多类歧义的动词总数约为 100 个。

表 1 显示了据此生成的初始链接库的基本统计数据，其中第 1 列的“SX”表示绝对单义，“MX”表示存在/拥有类单义，“MM”表示存在/拥有类多义情况。从中可以看出，目前关心的 3380 个动词词条中包含有效关联对的动词总数为 2197 个，占词条总数的 65%。它们反映了存在/拥有类动词在真实文本中的实际使用情况。同时，从句法树库中提取出这些关联对相关的句法树标注句子约 2.3 万句、50 万词，从 6 个月的人民日报标注库中提取出词语切分和词性标注句子约 21 万句、500 万词，形成了较大规模的以词汇关联对描述为中心组织的词汇语义计算基础数据。

在目前的初始链接库中，绝对单义动词（SE, SL,SH）又分别占了有效动词总数的 69%和关联对总数的 43%，平均每个动词包含约 37 个关联对，可以作为一个很好的语义标注研究出发点。

义项类别	动词数目	关联对数目
SE	593	23278
SH	564	18064
SL	367	14365
ME	210	14526
MH	269	27700
ML	108	11316
MM	86	19449
总计	2197	128698

表 1 初始链接知识库的基本数据

表 2 分析了 SE 和 SL 两大类绝对单义存在类动词的关联对频度分布情况。按照每个动词包含的关联对总数(RF)的不同，分成以下 4 类：1）RF∈[1,6)；2) RF∈[6,100)；3) RF∈[100,500)；4) RF>=500。从中可以看出，包含 100 个以上关联对的动词虽然只占了相应动词总数的 9%和 7%，但其覆盖的关联对比例则分别达到了 65%和 71%以上，平均每个动词分别覆盖 295 和 424 个关联对，显示了很强的信息集聚性，为进行深入的词汇语义计算打下了很好的基础。下面分别列出了两类中覆盖关联对数目大于 500 的动词词条：SE：出现、产生、形成、存在、提出、造成、建设；SL：开始、建立、组织、实行、中断。

	SE		SL	
频度类别	动词数目	关联对数目	动词数目	关联对数目
1	281	662	208	455
2	261	7559	135	3740
3	44	8496	19	3265
4	7	6561	5	6905
Tot	593	23278	367	14365

表 2 两类存在类动词的关联对频度分布数据

6.2 绝对单义动词关联对的参量锚定分析

在本文写作之时，我们初步完成了对 SE 和 SL 类的约 3.8 万关联对的参量锚定自动标注和人工校对工作。本节将以 SE 类为例对有关处理结果进行初步分析。

首先，利用语言学家对汉语存现句的大量研究成果，我们总结了以下启发式规则，通过开发自动参量锚定工具，实现了初步的句法语义链接：

1. (Syn=PO) && B-在|于 → L //述补结构 V
2. (Syn=ZW) && (时空词或方位结构) → L
3. (Syn=DZ|PO|ZW) → x // default
4. (Syn=ZZ) && (时空词或方位结构) → L
5. (Syn=ZZ) && 其他介词 → O
6. (Syn=DZ-A) && 自动词 → x
7. (Syn=DZ-H) && 他动词 → y
8. (Syn=PO) → y
9. (Syn=ZW) → x

其中规则 1-3 主要针对静态的存在状态描述动词，并对汉语中不同的存现句式进行了特殊处理。规则 4-9 则针对其他存在状态变化类动词，通过综合考虑词汇关联库和情境语义库可以提供的关联词句法功能位置，包括状语(ZZ)、定语(DZ-A)、定中结构中心语(DZ-H)、宾语(PO)、主语(ZW)，句法关系组合标记（如：介词、方位结构等）和关键动词的子范畴信息（如：自动词、他动词）等来确定合适的参量锚定标记。

通过使用启发式规则处理，我们分别对 SE 和 SL 类中的 72%和 76%的关联对实现了自动标注。将这些自动标注结果与最终的人工校对结果进行差异性比较，我们发现两者的信息差异率分别为 35%和 31%左右，其中绝大多数是自动标注工具没能处理的比较复杂的定中关系描述实例。两者对照分析，可以看出目前的基于简单启发式规则的自动参量锚定的处理精度已达到 90%以上。这表明目前的启发式规则处理是比较有效的。在此基础上，人工校对和标注速度可以达到 200 对/小

时，并且交叉校对的差异率在 4%以下，显示出较高的处理效率和较好的标注精度。

为了更深入分析不同存在类描述情境的参量锚定分布特点，我们把 SE 类动词进一步分成以下两类：1）SE1 类：主要描述静态的存在状态，其典型情境表达式为：exist(x,L)；2）SE2 类：主要描述动态的存在状态变化，其典型情境表达式为：DO(x,P(x,y)) CAUSE exist(y,L)。参量锚定的重点是解决其中动作主体 x、存在主体 y 和时空范畴 L 与不同句法功能位置的对应关系。

表 3 和表 4 分别列出了人工校对结果中这两类动词的整体句法语义链接分布数据。其中第 1 列为句法功能描述，第 2-6 列分别是相应的不同锚定参量的出现频度和分布比率。从中可以看出：1）在主宾语位置上，针对 SE1 类动词，参量锚定的主要矛盾是存在主体和时空范畴的区分问题；针对 SE2 类动词，则是动作主体和存在主体的区分问题，因为深层宾语的主题化（曹逢甫，2005）是汉语中的一种常见语言现象；2）在定语和中心语位置上，参量锚定的主要矛盾则是内部语义联系(x,y,L)和外部修饰联系(Q)的区分问题。

这些客观的定量分布数据为我们下一步改进目前的自动参量锚定算法指明了方向。例如：可以通过对时空范畴的细致分类并引入更多的词汇化信息，更好地区分出典型 L 参量；可以通过动作主体和存在主体的词汇语义分布差异区分 x 和 y 参量；可以通过总结典型的修饰中心名词表实现内部语义联系(x,y,L)和外部修饰联系(Q)的区分。

	x	y	L	O	Q
PO	1891/ 85%	0/ 0%	306/ 14%	4/ 0%	15/ 1%
ZW	2275/ 88%	0/ 0%	221/ 9%	52/ 2%	35/ 1%
DZ-A	1093/ 95%	0/ 0%	20/ 2%	3/ 0%	39/ 3%
DZ-H	1058/ 82%	0/ 0%	50/ 4%	6/ 1%	172/ 13%
ZZ	6/ 1%	0/ 0%	264/ 68%	73/ 19%	46/ 12%

表 3 SE1 类的句法语义链接分布数据

	x	y	L	O	Q
PO	23/ 0.4%	6042/ 98%	41/ 1%	9/ 0.2%	21/ 0.4%
ZW	2306/ 72%	732/ 23%	131/ 4%	40/ 1%	16/ 0%
DZ-A	178/ 8%	1705/ 73%	166/ 7%	36/ 2%	238/ 10%
DZ-H	195/ 5%	2124/ 59%	101/ 3%	58/ 2%	1134/ 31%
ZZ	41/	76/	599/	266/	25/

	4%	8%	60%	26%	2%

表 4 SE2 类的句法语义链接分布数据

7 分析与讨论

近年来，在英语方面开发完成了两个大规模的句法语义标注语料库：PropBank 和 FrameNet，它们在语义体系和构建方法上各有特色。

PropBank 的语义描述基础是 Levin(1993)提出的动词语义分类体系，其基本假设是不同的动词句法组合隐含着不同意义区别。Kipper(2000)对此进行了归纳总结，使用 Arg0-Arg5 等标记来表示各个动词的不同语义角色，形成 VerbNet 库。在此基础上，通过对英语宾州树库的各个句法树标注句子中的核心动词语义类和相应语义角色的人工分析和标注，构建完成了 Propbank。

FrameNet 的语义描述基础是 Fillmore(1982)提出的框架语义学理论，将一组意义相近的动词、名词和形容词等组织成框架，通过设计不同的框架元素，形成完整的框架意义解释。构建过程采用了框架驱动方法：针对每个框架，选择其中的目标词，从大规模语料库中提取若干典型例句，人工分析其中目标词的具体语义，并对符合框架意义的句子完成句法语义信息的人工标注。

另外，研究人员将类似的语义体系和构建方法移植到汉语上，也开发了汉语 Propbank(Xue and al. 2002)和 FrameNet(刘开瑛 2006)，但规模还不是很大。

本文提出的汉语句法语义链接库开发方案的主要特点则在于：1）在特定的词汇关联对上，同时描述它在真实文本中可能形成的句法组合关系和语义角色关系，在词汇层面上直接建立起词汇对的句法语义链接关系；2）建立以静态关系状态描述为基础，并融合与其变化发展密切相关的动态动作行为的情境语义 Schema，形成针对某个特定语义问题的完整情境语义描述体系；3）融合多种语言资源和知识获取工具，开发大规模的词汇关联库和配套的标注语料库，为深层次的词汇语义计算提供足够规模的置信度较高的基础数据；4）开发半自动的语义信息标注工具，从单义词对到多义词对，从关联对到真实文本句子，逐步扩大句法语义链接知识库的标注规模和信息容量，提高整体的知识库开发效率。

目前，我们在单义存在类动词关联对的参量锚定处理中已取得初步实验效果。在今后的研究中，我们将逐步开发词汇语义本体计算、多义词义项排歧等工具，并把处理对象扩展到拥有类动词，争取尽快完成存在/拥有类的完整知识库。

致谢

本项研究得到国家自然科学基金资助（项目号：60573185, 60520130299）。感谢李丽、杨乐、薛冬晗同学帮助完成了相关词汇对的语义信息标注和校对工作。

参考文献

Charles J. Fillmore, Charles Wooters, and Collin F. Baker, 2001. Building a Large Lexical Databank Which Provides Deep Semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation. Hong Kong*.

D. Gildea, and D. Jurafsky 2002 Automatic labeling of semantic roles. *Computational Linguistics*, 28(3): 245-288.

Paul Kingsbury, and Martha Palmer 2002. From Treebank to PropBank. *In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain*.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm, p. 111-137, Hanshin Publishing Co., Seoul, South Korea.*

Karin Kipper, Hoa Trang Dang, and Martha Palmer 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventh National Conference on Artificial Intelligence(AAAI-2000), Austin, TX, USA*

Beth Levin 1993. *English Verb Classes and Alternations A Preliminary Investigation*. MIT Press.

R. D. Van Valin, and R. J. Lapolla 1997 *Syntax: Structure, meaning and function*. Cambridge University Press.

Nianwen Xue, F. Chiou, and Martha Palmer 2002. Building a large-scale annotated Chinese corpus. In *Proc. of COLING-2002*.

曹逢甫著，王静译 2005 汉语的句子与子句结构．北京：北京语言大学出版社．

刘开瑛，由丽萍 2006 汉语框架语义知识库构建工程．曹右琦、孙茂松主编《中文信息处理前沿进展》，北京：清华大学出版社：p64-71．

苑春法，许伟，黄昌宁 1997 汉语语义关联网的研究．陈力为,袁琦主编：语言工程．北京：清华大学出版社，145-150．

俞士汶，朱学峰，王惠等 1998．现代汉语语法信息词典详解．北京：清华大学出版社．

张卫国，冀小军等 1994 现代汉语辞海．北京：人民中国出版社．

周强 2004．汉语句法树库标注体系．中文信息学报，18(4)：1-8．