文章编号：1003-0077(2017)05-0021-11

基于概念知识树的双宾短语分析

林子琦[1,2]，倪晚成[1]，赵美静[1]，杨一平[1]

(1. 中国科学院自动化研究所 综合信息系统研究中心，北京 100190；
2. 中国科学院大学，北京 100049)

摘 要：双宾短语是一种特殊的语言现象，为了使计算机能够理解并处理双宾短语，该文从语法和语义两个层面对双宾短语进行了分析，基于概念知识树知识表示模型建立了双宾短语的语义表达模型；并提出一种双宾短语分析算法，实现了从双宾短语到其语义表达模型的自动转换。双宾短语分析算法采用自顶向下和自底向上相结合的方法，自顶向下用于对双宾短语的语法成分进行划分，获得构成双宾短语的双宾动词成分、间接宾语成分和直接宾语成分；自底向上用于使用基于概念知识树的短语分析推理算法对双宾短语中的这三种成分分别进行分析，获得对应的语义表达；最后，利用三种成分的语义分析结果构建双宾短语完整的语义表达。该文从权威文献和语法词典中选取了 122 个双宾动词，对这些双宾动词构成的 209 个短语进行了分析，分析的正确率为 90.43%，证明了该文提出的双宾短语分析算法和语义表达模型的有效性。

关键词：双宾短语；概念知识树；语法分析；语义表达模型

中图分类号：TP391　　**文献标识码**：A

Parsing of Double-Object Phrases Based on Concept Knowledge Tree

LIN Ziqi[1,2], NI Wancheng[1], ZHAO Meijing[1], YANG Yiping[1]

(1. Integrated Information System Research Center, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: This paper analyzes the double-object phrase which is a special linguistic phenomenon from the syntactic and semantic perspective, and presents a semantic double-object expressive model based on Conceptual Knowledge Tree (CKT). Moreover, this paper proposes a method for analyzing the double-object phrases, which can automatically translate them into the semantic expressive model. It firstly, in a top-down style, classifies the syntactic parts of a double-object phrase into three parts - double-object verb, direct object and indirect object. And then, in a bottom-up style, it uses CKT to do inferences on these three parts and get their semantic expressions. Experiment on a dataset consisting of 122 double-object verbs and 209 phrases selected from authoritative literatures and grammar dictionaries reveals an accuracy 90.43%.

Key words: double-object phrases; concept knowledge tree; syntax parsing; semantic expressive model

1 引言

双宾短语是一种特殊的语言现象，现代汉语中的双宾语[1]是“某些动词能带两个宾语，一般是一个宾语指人，另一个宾语指物，如‘我问你一句话’。指人的一个(‘你’)靠近动词，叫作近宾语；指事物的一个(‘一句话’)离动词较远，叫作远宾语”。

语言学家对双宾短语的结构主要有两种观点[2]，第一种观点认为双宾短语应该分析为动宾短语带宾语；第二种观点认为双宾短语是谓语动词分别与两个宾语发生关系，两个宾语在同一层次，一个是近宾语，另一个是远宾语。相比于第一种观点，本文认为第二种观点更加清楚地刻画了两个宾语和谓语动词的关系，有利于计算机的分析和理解。因此，本文依据第二种观点对双宾短语进行分析。

收稿日期：2016-09-14　定稿日期：2017-03-13

目前语言学家对于双宾短语的研究很多,然而,在计算机领域中,专门针对双宾短语的研究相对较少。我们需要对双宾短语进行形式化的表达,从而让计算机能够处理乃至理解这种特殊的语言现象。如果不能正确地分析双宾短语,会导致很多后续的研究产生错误。例如,“转赠学校一批图书”表达的是将原本不属于学校的“一批图书”转赠给了学校;但是,如果“转赠学校一批图书”被错误分析为单宾短语,即“学校”和“一批图书”构成偏正短语做“转赠”的宾语,语义变成将属于“学校”的“一批图书”转赠了出去;并且,如果让计算机回答“这批图书被转赠给了谁”,那么计算机会找不到“转赠”的对象,导致分析出错。如果将“告诉他下雨了”分析成“他下雨了”构成主谓短语做“告诉”的宾语,从语法层面上看,上述分析是可以的,但是从语义上来说,人是不会下雨的,这样的分析结果是完全不合理的。

因此,对双宾短语进行正确合理地分析不仅是语言学家要研究的,也是计算机领域中不可忽视的问题。为此,本文研究了双宾短语的语义表达模型,并提出一种基于概念知识树的双宾短语分析算法,从而帮助计算机更加准确地理解双宾短语的语义。

本文第二部分介绍概念知识树知识表示模型;第三部分分析双宾短语的特征,并建立双宾短语概念库;第四部分对双宾短语分析算法的分析基础进行了介绍,即描述基于概念知识树建立的汉语语法知识树;第五部分详细阐述了本文提出的双宾短语分析算法;第六部分给出实验结果和分析;最后总结了本文的工作。

2 概念知识树知识表示模型

人类的知识需要经过编码,才能为计算机所理解,构建知识的表示模型是人工智能的基础[3]。在人工智能领域,很多知识表示方法被提出,如谓词逻辑[4-5]、产生式表示法[6]、框架[7]、语义网络[8-10]等。这些方法在某些特定领域取得了很好的效果。但是,这些方法在理论和实际的应用中都存在一定的局限性,如谓词逻辑难以表达不确定性知识和启发性知识;产生式表示法在处理复杂问题时,容易引起组合爆炸问题;框架不善于表达过程性知识;语义网络对知识的表示不能保证其不存在二义性。

通过研究和借鉴语义学、本体论、形式逻辑等学科对于知识表示和推理的理论,中国科学院自动化研究所综合信息系统研究中心构建了一种新的知识表示模型——概念知识树(concept knowledge tree,CKT)知识表示模型[11-13]。该模型分为概念和知识树两个层次:第一层是概念层,该层基于概念的语义表达模型,通过形式化的表达刻画语言的内涵,是知识表示模型的核心和基础;第二层是知识层,该层基于树结构的知识表达模型,具体如图1所示。

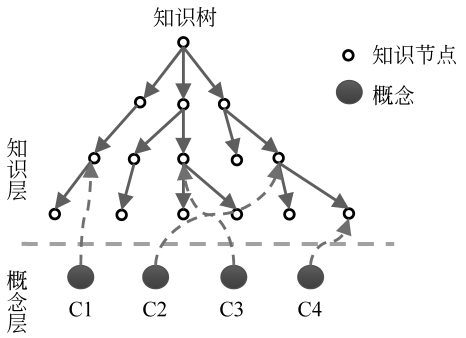

图1 概念知识树知识表示模型

2.1 基于概念的语义表达模型

概念知识树知识表示模型将概念[12]作为表达语义的基本单元,使用属性、关系和行为三个要素来对概念这一思维单元进行刻画,即在概念知识树知识表示模型中:

概念 = {概念名称,属性,关系,行为} (1)

其中属性是概念的基本特征,对概念起表示和区分的作用;关系是概念间的相互联系,概念的属性和行为因关系的存在得以继承和发展;行为是概念间的相互作用,体现了事物的运动特性,是概念发展变化的源动力。

概念又分为独立概念和复合概念,其中独立概念在语义上是不可再分的单元,例如,概念“猫”“花”等;而复合概念是由两个或两个以上的独立概念复合而成,如“巡航导弹”由概念“巡航”和概念“导弹”复合而成。

复合概念根据其复合方式划分为三种类型:语义约束,语义逻辑,语义状态,其中三种复合方式之间及其本身可以相互嵌套,从而形成复合概念。

语义约束主要用于表达由概念间修饰和属性约束关系构成的复合概念,其中一个概念是复合概念的核心概念,另一个概念是约束概念,用于修饰核心概念。语义约束可以用一个二元组来表示,如式(2)所示。

语义约束 =< 约束概念:核心概念 > (2)

例如,“红花”表示成“<红:花>”。

语义状态用于描述事件的语义复合结构,由谓

词概念、主体概念、客体概念和状态概念四个部分构成。其中，谓词概念用来表示事件的行为和动作等；主体概念和客体概念用来表示事件的主格和宾格；状态概念用来表示事件发生的条件、背景等。语义状态可以用一个四元组来表示，如式(3)所示。

语义状态＝[＜主体概念＞{状态概念}
谓词概念＜客体概念＞]　　(3)

例如，“他明天去北京”表示成“[＜他＞{明天}去＜北京＞]”。

语义逻辑用于描述通过连词等连接在一起的多个独立概念或复合概念之间的逻辑组合关系。语义逻辑由两部分组成，逻辑关系和概念列表。其中，逻辑关系分为六种：“逻辑与”“逻辑或”“逻辑表”“逻辑异或”“逻辑非”“逻辑蕴含”。“逻辑与”表示概念并列，“逻辑或”表示概念具有选择性，“逻辑表”表示若干概念的简单排列，“逻辑异或”表示概念的异同性，“逻辑非”表示概念否定，“逻辑蕴含”表示概念间的因果、递进关系。语义逻辑可以用一个二元组来描述，如式(4)所示。

语义逻辑 ＝ (逻辑类型，概念列表)　　(4)

例如，“我和你”表示成“(与，我 你)”。

2.2　基于知识树的知识表达模型

如图 1 所示，在概念知识树的知识表示模型中，以知识树的形式组织和存储领域知识。知识树以概念语义表达为基础，每个知识节点是以独立概念或者复合概念为意义的基本表达单元，并且每个知识节点上可以构建相应的属性、规则，从而详细地刻画知识节点的特性。知识节点之间的有向边表示知识节点间的父子关系或成员关系。

其中，对于知识节点间的共性只在父节点中保存，子节点可以继承并获取到这些信息，减少了信息存储的代价；而对于知识节点间的特性，可以在各个节点中按需刻画并扩展。

基于知识树的知识表达模型，从某一侧面对特定领域的知识进行切分，将切分好的知识表示成计算机可以存储和识别的数据，即构建对应的知识节点；然后，按照知识节点间的关系将知识节点组织成树状结构，从而更好地刻画知识间的共性和特性。知识树是表达知识的层次结构[11]，在各层中，知识节点与其相邻节点在粒度上保持一致；根节点层次最高，叶节点层次最低，层次越高，粒度越大；层次越低，粒度越小。

3　基于概念知识树的双宾短语语义表达模型构建

3.1　双宾短语的特征分析

(1) 双宾短语的语法特征

本文从《汉语动词用法词典》[14]、《现代汉语语法信息词典详解》[15]和文献[16]中，抽取出 122 个双宾动词和上述双宾动词构成的 209 个短语；其中，主要研究上述 122 个双宾动词构成的三种类型的双宾短语，三种类型的双宾短语如下所示：

① V＋N1＋N2

双宾动词后面带的两个宾语都是体词性的成分。例如，“送你一本书”“称他老三”；

② V1＋N1＋V2

双宾动词后面的近宾语是体词性成分的，远宾语是谓词性的。例如，“告诉他下雨了”；

③ V＋N1＋小句

双宾动词后面的近宾语是体词性成分的，远宾语是小句。例如，“告诉他明天我找他”。

通过分析发现，双宾短语的构成是一个动词后面带有两个宾语，近宾语称为间接宾语，远宾语称为直接宾语。虽然，有些语言学家认为像“吃了他三个苹果”不是双宾短语，“他三个苹果”是表示领属关系的偏正短语，与“吃”构成动宾结构；但是，文献[16-17]也从多个角度论证了“吃了他三个苹果”可以分析为双宾短语。

本文认为“吃了他三个苹果”可以分析为双宾短语，即将“他”分析为“吃”的间接宾语，“三个苹果”是“吃”的直接宾语。本文将此类短语分析为双宾短语本质上是为了更加全面地刻画其特征，为下一步的研究提供更加丰富的语义信息。

(2) 双宾短语的语义特征

分析发现，双宾短语的语义特征主要是通过双宾动词的语义体现，即双宾动词执行后会导致直接宾语与主语、间接宾语间的领属关系发生变化。其中，文献[18-19]认为在双宾句中，直接宾语所指对象在主语和间接宾语之间会发生转移。参考语言学家对双宾动词的分析，按照直接宾语所指对象在主语和间接宾语的转移方向，本文将双宾动词分为三类。

① “给予类”表示直接宾语从主语转移到间接宾语；例如，双宾短语“送你一本书”，其语义特征为

“书”从双宾动词“送”的发出者(主语)转移到了间接宾语“你”的手上;

② “获取类”表示直接宾语从间接宾语转移到主语;例如,“吃了他三个苹果”的语义特征为“三个苹果”从间接宾语“他”转移到了动作发出者(主语);

③ “转移不明类”表示直接宾语在主语和间接宾语间的转移方向不明确或不存在转移;例如,“掰你一块馒头”可以表示“掰了你的一块馒头”和“掰给你一块馒头”两种语义,可以表达两种转移方向;“称他老三”不存在直接宾语的转移。

3.2 双宾短语的语义表达模型

通过对双宾短语的分析,我们发现双宾短语可以使用“[谓词概念<客体概念>]”的动宾短语的语义表达模型来表示;但是,双宾短语与动宾短语相比,其谓语动词后面带有两个宾语。对于这种语义关系,我们基于概念知识树知识表示模型的“语义逻辑”来表示;其中,“语义逻辑”中的“逻辑表”表示若干概念的简单排列,可以用来表示双宾短语的两个宾语,即通过语义状态和语义逻辑的嵌套,构成双宾短语的语义表达模型。最终,双宾短语的语义表达模型如下:

[双宾动词概念<(表 间接宾语概念 直接宾语概念)>]

例如,“安排他两间房”对应的语义表达为“[安排<(表 他 <<两:间>:房>)>]”,如图 2 所示。

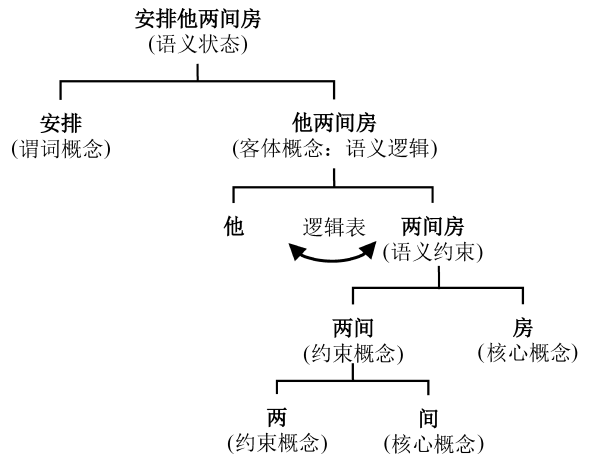

图 2 双宾短语语义表达模型实例

3.3 双宾短语概念库的构建

概念知识树知识表示模型以概念为表达语义的基本单元,以概念库作为语义计算和分析的基础。实验室现有概念库中包含九万多个中文概念(包含独立概念和复合概念),其中 76 828 个概念是基于《现代汉语词典》构建[20]的。双宾短语概念库是基于已有概念库构建的,其中主要分为以下两个部分。

(1) 基本概念的构建

在已有概念库的基础上,构建 122 个双宾动词对应的双宾动词概念;抽取出 209 个短语实例所包含的动、名词,并在概念库中构建缺失的动、名词相对应的概念。当然,对于 209 个短语所涵盖的各种词都可以构建相应的概念;但是,概念库本身已经涵盖大多数常见的概念,少部分概念的缺失并不影响目前双宾短语的分析,故对其余的概念暂时不构建,使用“缺失”这个概念来代替。

(2) 概念属性的构建

对短语实例中包含的名词概念,按照所表达语义的不同,对这些名词概念进行粗略的分类,主要分为以下几类:个人、集体名词(国家、连队)、机构(国务院、北京大学)、动物、植物、处所、抽象名词、物质名词、时间名词等;并对这些名词概念添加“类型”这个属性。

对双宾动词概念和短语实例中包含的动词概念添加“主体”“客体”“类型”这三种属性。其中,“主体”和“客体”分别填充动词概念能够带的主语和宾语的类型,这些类型对应于名词概念“类型”属性中的值;“类型”属性主要用来表征动词概念是否是双宾动词、是否能用于兼语句等。

最后,对双宾动词概念添加“方向”属性,即用该属性来表示双宾动词对直接宾语移动方向的刻画;其中,“方向”的属性值有三种:“给予类”“获取类”“转移不明类”。通过对双宾动词概念添加其特有的属性、行为等,可以将其语义特征更好地刻画出来,为下一步分析短语的语义提供有用的信息。

4 基于概念知识树的汉语语法知识树的构建

在双宾短语分析中,双宾短语概念库是双宾短语分析的基础,双宾短语的语义表达模型是对双宾短语的形式化表达,而语法正是双宾短语解析为形式化表达的桥梁。语法描述了语言中词、短语、句子等语言单位的组织规律。基于概念知识树语义表达模型,我们构建了基本的汉语语法知识树。

4.1 汉语语法知识树的基本结构

参考汉语语法书籍[21-23]对语法的划分,本文目前从词和短语两个层面来构建基本的汉语语法知识树:词包括名词、动词、形容词、副词、代词、介词、数

词、量词、数量词、连词、助词等；短语包括联合短语、主谓短语、同位短语、方位短语、数量短语、偏正短语、动宾短语、述补短语、介词短语[①]、连谓短语等。依据各类词、短语之间的关系，构成的汉语语法知识树的结构如图 3 所示。

在后续研究中，我们将继续加入句子等语言单位完善汉语语法知识树。

本文针对目前汉语语法知识树中的 11 种短语，构建了这些短语的基本语义表达模型，具体如表 1 所示。

表 1　短语的基本语义表达模型

短语类型	语义表达模型
联合短语	(与 概念 1 概念 2 …)
主谓短语	[<主体概念>谓词概念]
同位短语	<约束概念:所指概念>
方位短语	<约束概念:由方位词构成的概念>
数量短语	<数词概念:量词概念>
偏正短语	<约束概念:核心概念>
动宾短语	[谓词概念<客体概念>]
述补短语	[谓词概念{补语概念}]
介词短语	[介词概念<客体概念>]
连谓短语	(蕴含 谓词成分_1 谓词成分_2)
兼语短语	(蕴含 [谓词成分_1<兼语成分>][<兼语成分>谓词成分_2])

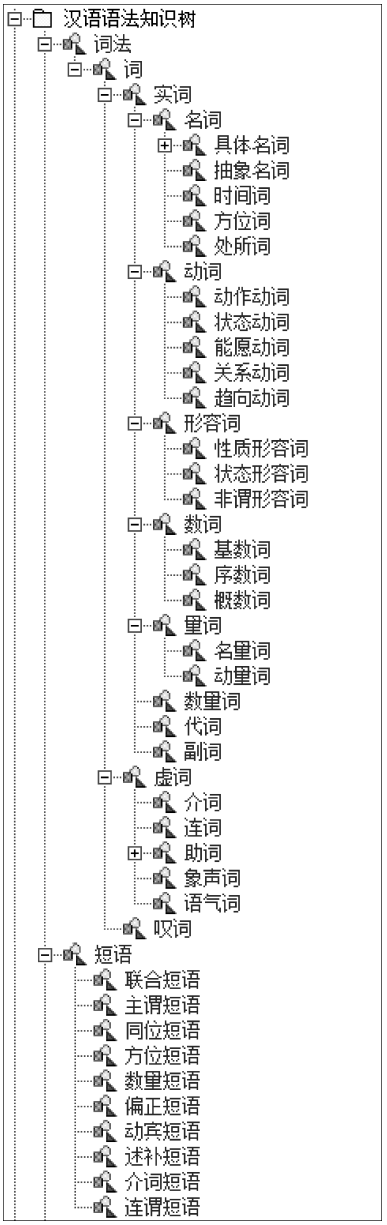

图 3　汉语语法知识树结构

4.2　基于汉语语法知识树的语法规则构建

4.2.1　汉语语法知识树的语法规则表示

在汉语语法知识树的基础上，我们在知识节点上构建了相应的语法规则。其中，规则的构建主要是从语法层面出发，即根据词序列中的词性、标志词(如“的”“地”“和”)等特征来形成规则；例如，在名词节点可以构建如下的规则：如果两个名词之间用“和”进行连接，则这两个名词形成联合短语，即“名词 1 和 名词 2”被识别成联合短语“(与 名词概念 1 名词概念 2)”。

知识节点上的规则采用 IF-THEN 的产生式规则来表达，例如，

```
If    (And (= (* input.pos* *cur*) m)(= (* input.pos* (*cur*+1)) q))
Then  (Set (*input* *cur*)(MeanBind (* input * *cur*)(* input * (*cur*+1))))
      (Set (* input.pos * *cur*) SL)
      (RemoveAt *input* (*cur*+1))
```
(5)

其中，“*变量名*”表示变量，“*input*”表示输入字符串经过概念映射得到的概念序列，“*input.pos*”表示概念序列对应的词性列表，“*cur*”表示当前分析程序处理到的概念序列的位置；式(5)为数词知识节点上的一条规则，表示的语

① 本文对介词短语主要关注介词短语中的介宾短语，其他介词短语待后面继续研究。

义如下：

如果：当前位置的概念词性为数词(m)，下一个位置的概念词性为量词(q)；

则：当前概念和后面的概念构成数量短语，并且进行如下操作，将当前概念变为当前概念和后面概念构成的语义约束、将当前概念的词性修改为数量短语(SL)、移除后面一个概念。

4.2.2 汉语语法知识树的语法规则规模

基于构建的汉语语法知识树，我们构建了一个包含196条规则的基础规则库，规则的具体分布情况如表2所示。

表2 汉语语法知识树上的规则分布情况

知识节点	规则数量	知识节点	规则数量
名词	18	介词	14
人名	2	助词	4
时间词	7	联合短语	6
方位词	1	主谓短语	6
动词	47	同位短语	2
趋向动词	5	方位短语	6
形容词	15	数量短语	6
状态形容词	1	偏正短语	7
数词	13	动宾短语	3
量词	1	述补短语	2
代词	19	介词短语	6
副词	5	总计	196

与一般的规则库不同，我们的规则是分散存储的，即按照规则的特性将规则建立在知识树的各个知识节点上。具体来说，按照规则前件第一个位置需要满足的概念词性对规则进行分类，将规则放置在该词性对应的词类或短语的知识节点上。通过这种存储方式，在对汉语文本进行分析时，可以根据待分析文本开头的概念有效地获取可能被激活的规则，提高分析程序的运行效率。

5 基于概念知识树的双宾短语分析算法

基于概念知识树的双宾短语分析算法，输入为待分析文本，输出为短语的语义表达，算法由预处理和双宾短语成分分析两个部分构成，具体流程如图4所示。

预处理首先对待分析文本进行分词和词性标注，得到有词性标注的词序列；然后，对词序列进行双宾短语边界识别、双宾短语词性矫正；最后，使用双宾短语成分分析对词序列进行分析，得到短语的语义表达。

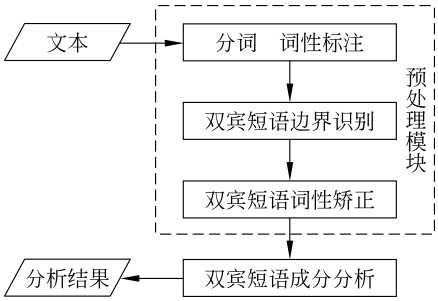

图4 基于概念知识树的双宾短语分析算法流程图

在预处理中，分词和词性标注使用中国科学院计算技术研究所的ICTCLAS 5.0；双宾短语的边界识别是通过识别词序列中的第一个双宾动词及离该动词最近的标点符号来得到双宾短语的范围；双宾短语词性矫正是利用规则将词性标注的一些明显错误进行纠正。双宾短语成分分析是算法的核心所在，将在5.1节中进行详述。

5.1 双宾短语成分分析

双宾短语成分分析采用的是自顶向下和自底向上相结合的分析方法；自顶向下是指在进行成分分析时，先对输入的短语序列进行成分划分得到双宾短语的双宾动词成分、间接宾语成分、直接宾语成分；然后，再对各个成分使用基于概念知识树的短语分析推理算法进行分析；最后，对各成分的语义表达进行消歧等操作后，将各成分分析结果进行组装，得到整个短语的语义表达，双宾短语成分分析算法流程如图5所示，主要步骤如下。

(1) 概念映射

双宾短语成分分析算法先对输入的词序列进行概念映射获取对应的概念序列，即找到词序列中每个词在概念库中对应的概念。由于一个词可能会对应多个概念，因此，用“词”加下划线和编号来表示该“词”所对应的概念。例如，“阿公”这个词分别对应三个概念：①丈夫的父亲；②祖父；③尊称老年男子，则概念库中分别用“阿公_1”、“阿公_2”、“阿公_3”来表示“阿公”代表的三个概念。目前，短语成分分析输入的是短语分词、词性标注后的词序列，文本长度较短；因此，暂时不考虑概念消歧的问题。

(2) 双宾动词成分的识别

虽然经过预处理后得到的词序列是以双宾动词

开头的，但是双宾动词后面有可能带有助词、介词等成分(“了”“给”等)，因而需要对经过概念映射得到的概念序列进行分析，识别双宾动词概念后面的附加成分，得到完整的双宾动词成分。例如，对概念序列“吃_2 了_7 他_1 三_1 个_1 苹果_1”进行双宾动词成分识别，得到的双宾动词成分为“吃_2 了_7”；

(3) 间接宾语成分的识别

将双宾动词成分后面的概念序列和间接宾语的模板进行匹配，如果匹配成功，则匹配成功的部分识别为间接宾语成分；反之，则将双宾动词成分后的第一个名词性概念识别为间接宾语成分；其中，间接宾语模板是统计收集到的 209 个短语实例的词性序列得到的。例如，“数词 量词 名词”(两 个 人)为统计得到的一个间接宾语模板。

(4) 单宾短语的判别

在对直接宾语的成分进行划分之前，检测前面的间接宾语成分是否一定会与后面的部分构成单宾语，判断标准是如果间接宾语的后面直接跟“是”“的”等标志性的词，则判定为单宾短语转到步骤(5)单宾短语的分析；否则，转到步骤(6)。

(5) 单宾短语的分析

通过步骤(4)词序列已被判定为单宾短语，则将整个词序列，即由步骤(1)得到的整个短语的概念序列，使用步骤(7)的基于概念知识树的短语分析推理算法进行分析，分析得到的语义表达即为单宾短语的分析结果，输出分析结果算法结束。

(6) 直接宾语成分的识别

通过步骤(4)词序列被判定为不是单宾短语，则继续进行双宾短语的分析，将间接宾语之后的部分识别为直接宾语。

(7) 基于概念知识树的短语分析推理算法

对步骤(2)、(3)、(6)得到的双宾动词成分、间接宾语成分和直接宾语成分，使用基于概念知识树的短语分析推理算法进行分析，得到其对应的语义表达，具体的算法见算法 1、算法 2。

(8) 语义消歧

虽然经过上述步骤的分析得到了三个成分的语义表达，但是分析算法中的规则主要是短语间的结构信息，并不能消除语义上的歧义。因此，需要利用概念的语义信息来消除歧义，从而得到正确的语义表达。其中，语义消歧主要分为以下两个部分。

① 直接宾语的消歧：主要利用概念间的属性来进行消歧，例如，利用动词概念的“客体”属性和名词概念的“类型”属性可以判断这个动词概念和名词概念能否构成动宾关系。

② 兼语短语和双宾短语的消歧：当待分析短语中的双宾动词概念能够形成兼语句时，需要检测间接宾语成分和直接宾语成分是否构成主谓关系，如果构成主谓关系，则识别为兼语短语；反之识别成双宾短语；其中，利用动词概念的“主体”属性和名词概念的“类型”属性可以判断这个动词概念和名词概念能否构成主谓关系。例如，概念序列“发展_1/v 农业_1/n”会激活两条规则，从而得到偏正短语“<发展_1:农业_1>”和动宾短语“[发展_1<农业_1>]”两种推理结果；程序通过查询概念“发展_1”的“客体”属性为“机构、抽象名词”，“农业_1”的“类型”属性为“抽象名词”，说明概念“农业_1”与“发展_1”可以带的宾语类型一致，因此将上述概念序列识别成动宾短语。兼语短语的检测方法与上述例子类似，即检测直接宾语成分形成的复合概念中的“主体”属性是否包含间接宾语成分形成的复合概念中的“类型”属性，即可判断两种成分是否构成主谓关系。

(9) 短语语义表达的构造

根据语义消歧得到的短语类型对步骤(7)得到的各部分语义表达进行组装，构造出完整的短语语义表达，即按照“兼语短语”或“双宾短语”中各个成分间的关系，构造完整的短语语义表达。例如，对短语“补充连队三名新兵”经过步骤(1)～(7)的处理后(不进入步骤(5))，分别得到其各部分的语义表达，即双宾动词成分的语义表达“补充_1”、间接宾语成分的语义表达“连队_1”、直接宾语成分的语义表达“<<三_1:名_8>:新兵_1>”；再经过步骤(8)的语义消歧，程序判定上述短语为双宾短语，根据双宾短语的语义表达模型构造出完整的短语语义表达为“[补充_1<(表 连队_1 <<三_1:名_8>:新兵_1>)>]”。

5.1.1　基于概念知识树的短语分析推理算法

基于概念知识树的短语分析推理算法主要利用汉语语法知识树上的规则对短语对应的概念序列进行分析，从而获得短语的语义表达；算法的核心是基于规则的推理算法，即基于推理树的推理算法(算法 2)。

基于概念知识树的短语分析推理算法是在推理树上进行的，输入为汉语语法知识树、概念序列、推理位置，输出为概念序列的语义表达。其中，汉语语法知识树是推理树进行推理的知识和规则的来源；概念序列是词序列经过概念映射得到的；推理位置

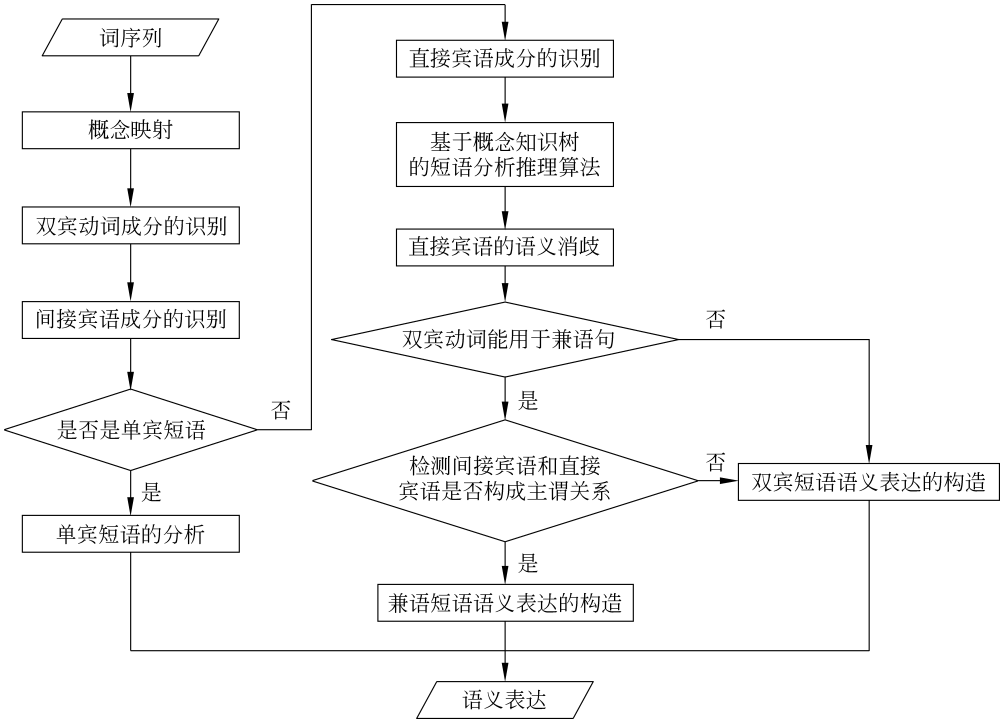

图5 双宾短语成分分析算法流程图

给定推理是从概念序列的哪个位置开始。算法的伪代码如下：

算法1 基于概念知识树的短语分析推理算法

```
1. Reason (knowtree, conceptlist, pos)
2.  reason_path = [], reason_result = [];
3.  rt = ReasonTree();
4.  rt.rt_Reason (knowtree, conceptlist, pos, reason_
                path, reason_result);
5.  WHILE((reason_result==[]) and
            knowtree.HasParent(conceptlist)) DO
6.      conceptlist = knowtree.findParent (con-
ceptlist);
7.      rt.rt_Reason(rt, knowtree, conceptlist, pos,
                reason_path, reason_result);
8.  END WHILE
9.  return reason_result
```

其中，第2、3行代码是对推理路径、推理结果和推理树进行初始化；

第4行代码表示使用基于推理树的推理算法对输入的概念序列进行分析，获得其对应的语义表达，基于推理树的推理算法见算法2；

第5到第8行代码表示当推理失败没有获得推理结果时，如果当前概念序列中有概念的词性在汉语语法知识树中有父节点的话，那么就将这些概念的词性变为父节点的词性，重新进行推理，即子知识节点可以爬升到父知识节点继承父节点的规则和语义信息；例如，汉语语法知识树中趋向动词的父节点是动词，当概念词性是趋向动词时，如果推理不成功，那么该概念节点的词性就爬升到动词这个父节点，从而利用动词节点的规则和语义信息继续进行推理。

5.1.2 基于推理树的推理算法

基于推理树的推理算法是基于规则的深度优先推理算法，输入为汉语语法知识树、概念序列、推理位置、推理路径、推理结果，输出为推理结果，即概念序列可能的语义表达。基于推理树的推理算法伪代码如下：

算法2 基于推理树的推理算法

```
1.rt_Reason(knowtree, conceptlist, pos, reason_path,
            reason_result)
2.  WHILE( pos<len(conceptlist) )
3.      topNodes = getRules (conceptlist [pos],
knowtree)
4.      curPath = getReasonPath();
5.      checkPath(curPath, reason_path);
6.      FOR i = 0:len(topNodes)-1
7.          rn_Reason(topNodes[i], conceptlist,
output);
8.      END FOR
9.      FOR i = 0:len(output)-1
10.         IF len(output[i])>1 THEN
11.             rt_Reason(knowtree, output[i],
```

```
                pos-maxRuleLen + 1, reason_
                path,reason_result);
12.             rt_Reason(knowtree, output[i],
                pos, reason_path, reason_re-
                sult);
13.         ELSE
14.             reason_result.append(output
[i]);
15.         END IF
16.     END FOR
17.     pos++;
18. END WHILE
```

其中，第 2 行代码表示当推理的位置未到达概念序列的末尾时，继续推理（推理位置从 0 开始计数）；

第 3 行代码表示获取当前进行推理的概念可能满足的规则，即利用概念的词性在汉语语法知识树上得到对应知识节点中的规则，并将这些规则作为推理树进行推理的顶层节点；

第 4～5 行代码表示获取当前推理路径，如果当前推理路径已经推理过，则函数返回，推理结束；否则，将保存当前推理路径，继续推理；

第 6～8 行代码表示对推理树的顶层节点进行推理，即检验当前输入的概念序列是否满足规则，如果满足规则，则执行规则后件得到推理结果；

第 9～17 行代码表示如果推理成功，则对每个推理结果进行两次推理，即将推理位置向前回溯一定距离（最大规则长度－1）或保持当前推理位置不变，然后，递归调用基于推理树的推理算法；反之，推理位置加 1，继续推理；

第 14 行代码表示当推理结果的长度为 1 时，表明已获得对应的语义表达，保存推理结果。

基于推理树的推理算法主要是深度优先的推理算法，采用递归的方式进行推理，并且记录推理路径，避免了重复推理。

例如，对短语“一块馒头”进行预处理、概念映射等处理后得到的概念序列“一_1/6 块_2/7 馒头_1/0”使用基于推理树的推理算法进行推理（概念使用数字来表示词性，例如名词：0、数词：6、量词：7、偏正短语：21、数量短语：25 等），推理过程如下：

①　一_1 块_2 馒头_1（0：6 7 0）

②　　<一_1：块_2> 馒头_1（0：25 0）

③　　<<一_1：块_2>：馒头_1>（推理结果）

④　　<一_1：块_2> 馒头_1（1：25 0）

⑤　一_1 块_2 馒头_1（1：6 7 0）

⑥　　一_1 <块_2：馒头_1>（0：6 21）

⑦　　一_1 <块_2：馒头_1>（1：6 21）

⑧　一_1 块_2 馒头_1（2：6 7 0）

其中，每一行表示当前推理时输入的概念序列，括号中表示的是“（推理位置：概念序列的词性）”；

第①行：从起始的“数词 量词 名词”组成的概念序列开始推理，取出汉语语法知识树上的“数词”节点中的规则，“数词 量词”激活规则被识别为数量短语，即第②行为推理结果；

第②行：第 1 行推理完成后，将推理位置向前回溯一定距离（最大规则长度－1）或保持当前推理位置不变，然后，递归调用基于推理树的推理算法，在此由于推理位置是 0 不能再向前回溯，故保持推理位置不变，从 0 开始推理，即对第②行进行推理，取出“数量短语”节点中的规则，“数量短语 名词”激活规则被识别为偏正短语，即第③行的推理结果；

第③行：第③行待推理的概念序列长度 1，说明已获得推理结果，推理位置加 1；

第④行：第④行与第②行属于同一层次的推理，但由于从位置 1（“馒头_1”）开始，没有可以满足的规则，故没有推理结果；

第⑤～⑧行的推理过程类似，通过推理路径可以观察到递归调用基于推理树的推理算法的层次，即概念序列的词性一致的为同一层次的推理，例如，①、⑤、⑧均为最外层次的推理。

由上述推理过程，可以发现基于推理树的推理算法是深度优先的方法，因而上述推理实例在第③行就得到了推理结果；推理路径为：[0：6 7 0；0：25 0；1：25 0；1：6 7 0；0：6 21；1：6 21；2：6 7 0]，通过比较当前推理路径是否在推理路径中出现过即可避免冗余的推理。

6　实验与分析

6.1　实验设置

实验数据是从《汉语动词用法词典》[14]、《现代汉语语法信息词典详解》[15]和文献[16]中，抽取出的 122 个双宾动词，及其构成的 209 个短语，数据的分布情况如表 3 所示。

表3 待分析短语集合数据分布情况

短语类型	数量	短语长度	字数
双宾短语	136	最大长度	14
兼语短语	70	最小长度	4
单宾短语	3	平均长度	6.79
总计	**209**	**总字数**	**1 419**

实验中使用的操作系统为 Windows 7(旗舰版64位)、处理器为 AMD A8-6500 3.50GHz、内存4GB、硬盘1TB、分词和词性标注工具为 ICTCLAS 5.0、数据库为 Microsoft SQL Server 2005。

6.2 实验结果

对209个短语实例使用基于概念知识树的双宾短语分析算法进行分析,将得到的语义表达进行分析,识别的正确率为90.43%,错误率为9.57%,具体的实验结果如表4所示。

表4 短语分析实验结果

短语类型	数量	识别正确	识别错误
双宾短语	136	127	9
兼语短语	70	60	10
单宾短语	3	2	1
总计	**209**	**189**	**20**

图6给出了分析得到的部分结果;由于一个词可能会对应多个概念,因此用"词"加下划线和编号来表示该"词"所对应的概念;每个实例的第一行为待分析短语,第二行为分词和词性标注结果,第三行为短语分析结果。其中,前两个实例被分析为双宾短语,第三个实例被分析为单宾短语,最后一个实例被分析为兼语短语。实例"派他去西藏工作"被分析为"派他"和"他去西藏工作"构成的兼语短语,"去西藏工作"被分析为连谓短语。

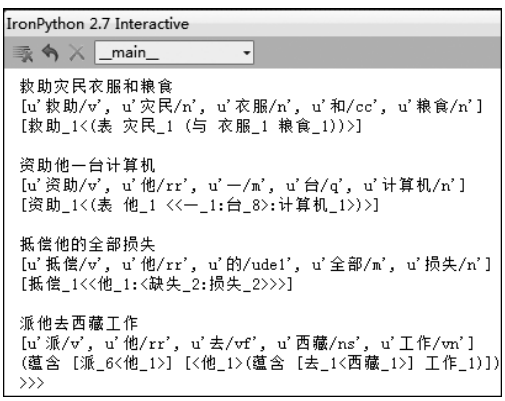

图6 双宾短语分析算法部分分析结果

图7给出了由于子节点"趋向动词"的规则不能推导出结果,从而爬升到其父节点"动词"继承父节点规则得到的分析结果。其中,"趋向动词"节点中不包含动词和名词构成动宾短语的规则。

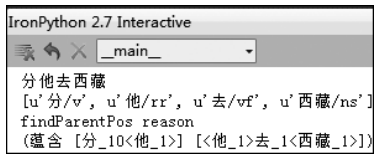

图7 子节点爬升到父节点的分析实例

6.3 实验分析

表5针对每个实例的错误原因进行了分析(括号中为矫正结果)。其中,对分词和词性标注错误进行矫正后,有十个实例能得到正确的分析结果;而剩下的两个实例在进行词性矫正后,还需要进行消歧才能得到正确的分析结果。

表5 短语分析错误类型统计

错误类型	数量	实例
分词错误	7	参加会议(参加 会议)月工资(月 工资)
词性标注错误	5	为/p(为/v) 给/v(给/p)
规则未覆盖	3	连谓短语、时间词等规则覆盖不全
歧义	5	"生活费用" 结构上可以构成动宾或偏正

在导致短语分析错误的四种原因中,分词错误和词性标注错误对分析结果的影响很大,如果要解决这类错误,则需要改善分词和词性标注的性能;对于规则未覆盖导致的错误,说明对规则的总结不到位,需要向规则库合理地增加规则,从而提高规则的覆盖度,并保证规则数量不会爆炸式地增长;前面几种错误主要是语言结构层面上的问题,而歧义问题是语义层面上的问题,同时也是最难解决的问题,需要添加语义信息辅助识别。

例如,"吃了他三个苹果"中由于在"吃"对应的概念中添加了其"方向"属性为"获取类",因而可以得到在发出"吃"这个动作前,"他"和"三个苹果"构成领属关系;然而,如果仅从结构上来进行辨析是分辨不了"他"和"三个苹果"是否构成领属关系,需要利用文献[24-26]中所述的语义信息和特征来进行

判别。

当构成偏正短语的两个成分之间没有“的”这种标志性的词时，要判别是否是偏正短语是很困难的，比如“生活费用”这种动词作定语的粘合式偏正结构[27]，就需要利用语义信息来进行判别。

7　总结与展望

为了准确地将汉语中的双宾短语表达为计算机可理解的语义表达模型，我们基于概念知识树构建了双宾短语的语义表达模型，并提出了基于概念知识树的双宾短语分析算法。实验证明本文提出的基于概念知识树的双宾短语分析算法是有效的。本文提出的双宾短语分析算法具有如下特点：

(1) 自顶向下和自底向上相结合的深度优先分析算法；

(2) 对当前概念可能满足的规则集，搜索的时间复杂度为 O(1)；

(3) 分析算法是利用汉语语法知识树和概念库中的信息进行推理；

(4) 概念知识树的这种求同存异的知识构建方式，使得汉语语法知识树上的知识节点可以存储自身特有的规则和语义信息，并且通过继承父节点的规则和语义信息，保证了信息的丰富性和完整性，同时减少了冗余信息的存储；

(5) 分析算法在利用短语的结构信息进行分析的基础上，利用概念的语义信息进行消歧，从而将语言的结构信息和语义信息相结合，增强了短语分析算法的分析能力。

本文构建了双宾短语的概念库和汉语语法知识树，并且有一定规模的概念库作为双宾短语分析的基础；然而，要想全面地实现双宾短语的分析，现有概念库和知识树中存储的规则和语义信息还远远不够，在后续的研究中，我们会进一步完善汉语语法知识树中的规则库，将更多的知识加入到概念库和知识树中。另一方面，语言学家对众多的语言现象进行了很深入和细致的研究，并且提出了很多用于分析语言的方法和特征；然而，这些方法和特征更多的是偏向于让人来使用的；因此，如何将这些有用的知识转化为计算机可利用，甚至可理解的信息，也是我们后续研究中需要继续思考的问题和努力的目标。

参考文献

[1] 中国社会科学院语言研究所词典编辑室. 现代汉语词典[M]. 第 5 版. 北京：商务印书馆. 2005.

[2] 何莉芳. 现代汉语双宾语句研究历程及分歧综述 [J]. 湖北广播电视大学学报，2010，30(7)：111-112.

[3] Bench-Capon T J M. Knowledge representation：an approach to artificial intelligence[M]. New York：Academic Press，1991.

[4] Cat B D，Bogaerts B，Bruynooghe M，et al. Predicate Logic as a Modelling Language：The IDP System [C]//Proceedings of the Computer Science，2014.

[5] Van Emden M H，Kowalski R A. The semantics of predicate logic as a programming language[J]. Journal of the Acm，1976，23(4)：733-42.

[6] Poria S，Cambria E，Ku L W，Chen G，et al. A Rule-Based Approach to Aspect Extraction from Product Reviews[C]//Proceedings of the 2nd Workshop on Natural Language Processing for Social Media. Dublin，2014：28-37.

[7] Minsky M. A framework for representing knowledge [J]. Readings in Cognitive Science，1974，8(76)：156-89.

[8] Nosek J T，Roth I. A comparison of formal knowledge representation schemes as communication tools：predicate logic vs semantic network [J]. International Journal of Man-Machine Studies，1990，33(2)：227-39.

[9] Havasi C，Speer R，Alonso J B. ConceptNet 3：a Flexible，Multilingual Semantic Network for Common Sense Knowledge[C]//Proceedings of the 3rd Recent Advances in Natural Language Processing. Philadelphia，2007：27-29.

[10] Niemann H，Sagerer G F，Schroder S，et al. Ernest：a semantic network system for pattern understanding [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，1990，12(9)：883-905.

[11] 于海涛，高一波，杨一平. 基于知识树的领域知识组织和应用 [J]. 计算机应用研究，2008，25(11)：3246-3248.

[12] 高一波. 一种基于概念的知识表达体系 [J]. 微电子学与计算机，2004，21(09)：71-74.

[13] 卢朋. 基于语义的现代汉语介词理解研究 [D]. 中国科学院自动化研究所博士学位论文，2008.

[14] 孟琮，郑怀德，孟庆海，等. 汉语动词用法词典 [M]. 北京：商务印书馆，1999.

[15] 俞士汶，朱德熙，陆俭明，等. 现代汉语语法信息词典详解[M]. 第 2 版. 北京：清华大学出版社，2002.

[16] 陆俭明. 再谈“吃了他三个苹果”一类结构的性质 [J]. 中国语文，2002(4)：317-325.

[17] 吴坤. 浅谈“吃了他三个苹果”结构 [J]. 青年文学家，2016(15)：138-9.

[18] 古川裕. 谈现象句与双宾语句的认知特点 [J]. 汉语学习，1997(01)：20-3.

(下转第 49 页)

吕国英(1964—)，硕士，副教授，硕士生导师，主要研究领域为自然语言处理。
E-mail：english@ sxu .edu .cn

李茹(1963—)，博士，教授，博士生导师，主要研究领域为自然语言处理。
E-mail：lir@ sxu .edu .cn

苏娜(1989—)，硕士研究生，主要研究领域为中文信息处理。
E-mail：374286185@ qq .com

(上接第 31 页)

[19] 李敏 . 双宾动词的词汇语义和双宾句式语义的互动 [J]. 世界汉语教学，2006(4)：55-66 .
[20] 赵美静 . 汉语词典义的语义理解研究 [D]. 中国科学院自动化研究所博士学位论文，2015 .
[21] 黄伯荣，廖序东 . 现代汉语：下册 [M]. 北京：高等教育出版社，2007 .
[22] 李德津，程美珍，金德厚，等 . 外国人实用汉语语法 [M]. 北京：北京语言大学出版社，2008 .
[23] 刘月华，潘文娱，故韡 . 实用现代汉语语法 [M]. 北京：商务印书馆，2001 .
[24] 田英华 . 表领属的“名＋名”偏正结构词的内部语义关系 [J]. 阜阳师范学院学报(社会科学版)，2005(5)：32-35 .
[25] 李宇明 . 领属关系与双宾句分析 [J]. 语言教学与研究，1996(3)：63-74 .
[26] 司富珍 . 双宾结构中的领属关系 [J]. 外国语文研究，2015(3)：2-11 .
[27] 潘国英 . 论动词作定语的两种结构形式 [J]. 湖州师范学院学报，2001，23(4)：59-63 .

林子琦(1991—)，博士研究生，主要研究领域为语义信息处理、问答系统。
E-mail：linziqi2013@ ia .ac .cn

赵美静(1985—)，助理研究员，博士，主要研究领域为语义信息处理、知识表示与应用、大数据处理等。
E-mail：meijing .zhao@ ia .ac .cn

倪晚成(1978—)，通信作者，高级工程师，博士，主要研究领域为知识表示与知识服务系统、资源共享与优化、大数据处理等。
E-mail：wancheng .ni@ ia .ac .cn