面向对象有标复句本体建模 *

肖 升[1,2], 胡金柱[1], 姚双云[1], 吴锋文[1]
(1. 华中师范大学 计算机科学系, 武汉 430079; 2. 湖南省第一师范学校 信息技术系, 长沙 410002)

摘 要: 基于面向对象方法为带标志构件的现代汉语复句子类(有标复句)建立本体模型,奠定中文信息处理复句层级的研究基础。在原有成果的基础上,利用关系标志与分句间的联系对标志构件进行句法分析,并在本体构造方法框架的指导下,用 UML 语言构造有标复句领域相关概念的本体模型。与已有成果相比,改进的模型能更精确、更深入地描写有标复句的特征。

关键词: 本体; 有标复句; 面向对象; 建模
中图分类号: TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)02-0552-03
doi:10.3969/j.issn.1001-3695.2010.02.041

Object-orient ontology modeling for tag complex sentence

XIAO Sheng[1,2], HU Jin-zhu[1], YAO Shuang-yun[1], WU Feng-wen[1]
(1. *Dept. of Computer Science, Central China Normal University, Wuhan* 430079, *China*; 2. *Dept. of Information Technology, Hunan First Normal College, Changsha* 410002, *China*)

Abstract: Established ontology model for modern Chinese complex sentence with tag components (tag complex sentence) based on object-orient method, lied a foundation for study of complex sentence stage on Chinese information dealing. This paper took syntactic analysis for tag components using the relationship between the link of clause and tag based on the original results. And also established an ontology model for the related concept of the field of complex sentence based on the methodological framework of ontology and UML. The improved model can describe the feature of tag complex sentence more precisely and more profoundly compare with the old model.

Key words: ontology; tag complex sentence; object-orient; modeling

0 引言

本文的主要研究对象是现代汉语有标复句,它是现代汉语复句的一个子类,是使用了关系标志,形成了特定句式的复句[1]。之所以选择有标复句作为研究对象,看中的正是此类复句语表上出现的标志。这一标志是一种客观存在并且可以成为客观标准的形式实体,它的语法功能是连接分句,并标志其相互关系[1];它对计算机识别、处理句内短语起着关键作用,并可以从深层控制分句间的逻辑关系[2]。因此,选择有标复句作为中文信息处理复句层级上的研究对象是可取的。除关系标志带来的优势外,有标复句还具有良好的概念层次结构,而文献[3,4]的研究已证明本体在具有良好概念层次结构的领域有着广泛的应用前景,因此,将本体和本体理论应用到中文信息处理复句层级上是可行且合适的。但目前这方面的成果并不多,主要体现在文献[5,6]中。其中文献[5]分析了复句的静态结构及它与小句的关系,并用面向对象本体建模方法构造了小句元模型,且在此基础上初步探讨了复句静态本体模型。文献[6]则从本体基本建模元语出发构造了复句本体模型。虽然这些成果无疑将有助于复句的本体研究,但它们都未能深入研究关系标志以及分句间的联系,而这些恰好在有标复句中体现得比较直观,因此,本课题的一个重要意义在于,以有标复句和关系标志为突破口,研究本体在复句句法分析领域的深层应用。

1 本体及相关概念

近十年来,虽然本体研究的分支日趋完善,但与本体相关的概念和术语却不尽一致,因此,在研究有标复句本体构造之前,有必要统一以下本体及相关概念的定义。

定义 1 本体论(ontology)。它特指哲学的一个分支学科,对客观存在的一切事物进行系统的解释或说明,借此来抽象客观事实的本质及其相互间的关联。

定义 2 概念化(conceptualization)。它指某一概念系统所蕴涵的语义结构,是对某一事实结构的一组非正式的约束规则[7]。

定义 3 本体(ontology)。是对于概念化的明确表达[8]。

定义 4 本体构造(ontology tectonic)。本体捕获、本体描述、已有本体集成及对新生本体模型评估的一系列过程[9]。

定义 5 本体理论(ontological theory)。它是一个逻辑理论,用于说明一系列词汇的特定含义。为达成描述概念化范畴的目的,使用一系列逻辑语言来表达,但此语言必须基于其本

收稿日期: 2009-07-06; **修回日期**: 2009-08-23 **基金项目**: 国家教育部重点研究基地重大研究项目(07JJD740063); 国家重点实验室开放研究基金资助项目(SKLSE04-018); 湖南省教育"十一·五"规划重点课题(XJK06AZC010); 湖南省第一师范学院科研课题(XYS09N04)

作者简介: 肖升(1980-), 男, 湖南武冈人, 博士研究生, 主要研究方向为中文信息处理(xiaosheng@mail.ccnu.edu.cn); 胡金柱(1947-), 男, 湖北宜城人, 教授, 博导, 主要研究方向为中文信息处理; 姚双云(1974-), 男, 湖南邵阳人, 副教授, 硕导, 主要研究方向为中文信息处理; 吴锋文(1982-), 男, 湖北麻城人, 博士研究生, 主要研究方向为中文信息处理.

体约定的限制,从逻辑语言中找出适当的特定模型来说明概念化范畴的特定含义。

文中所涉及到的相关概念均以上述定义为准。

2　有标复句本体构造

2.1　本体捕获

本体构造的第一步是本体捕获,即在领域专业知识的配合下确定重要的概念和关系,给出它们的精确定义,并确定其他相关术语[9];具体到本课题,就是要对有标复句进行句法分析,从中挖掘重要的概念和关系,并选取 Perez 等人定义的五个本体元语(属性、关系、函数、公理、实例)来构造本体。

根据文献[10]的观点,这种挖掘应该作用于两个方面,一方面是有标复句的实义构件,另一方面是有标复句的标志构件。文献[5,6]对复句实义构件的挖掘是扎实且深入的,考虑到有标复句是复句的子类,它们的成果对有标复句的本体捕获肯定是有价值的。但它们没有挖掘能够体现有标复句特点的标志构件,也没有分析标志构件与实义构件间的关系,因此它们的成果对有标复句的本体捕获而言是不完善的,而如何将其完善正是本文考虑的重点。

标志构件的核心构件是关系标志,分析关系标志时,除了文章前言部分提到的关系标志的语法功能外,还应重点分析关系标志的分类和语表形式。

文献[2]认为,关系标志应该分为篇章关系标志和分句关系标志,理由是,虽然有些关系标志位于某一分句中,但所表示的关系却超出了所位居的复句,且往往与前面的句子或段落甚至篇章发生关系,因此应该与连接分句的分句关系标志区分开来,称为篇章关系标志。

分析语表形式本质上就是分析关系词语,它包括分析关系词语的范围,语法单位大小、类别、充当成分和是否是准标等属性。

语法学中通常认为关系词语包括四类:a)句间连词,它们通常连接分句,不充当句子成分,如“因为、所以、虽然、但是”等;b)关联副词,它们既起关联作用,又充当句子状语,如“就、又、也、还”等;c)助词“的话”,它表示假设语气,总出现在假设分句末尾,标明分句间的假设结果关系;d)超词形式,它们本身已不是一个词,如“如果说、若不是、不但不、总而言之”等[1]。

从上述关系词语的范围可知,关系词语语法单位所处的级是不固定的,可能是词,也可能是比词大的单位。例如“因为”“所以”是词,而“与其说”“不如说”则是比词大的单位。正因为关系词语语法单位大小不固定,所以关系词语在词类系统中的类别也不固定,可以是连词、副词、助词,还可以是多种类别的组合。例如“因为”“所以”是连词,“都”“就”是副词,“的话”是助词,“还是”是副词 + 判断动词。关系词语类别的不固定又导致关系词语在句中充当成分的差异,有的仅充当标明分句关系的语法成分,有的却能在充当语法成分的同时兼做句子成分。例如在“无论 p,都 q”的句式中,“无论”只充当标明关系的语法成分,“都”却既起关联作用又兼做状语。虽然关系词语是标志关系的语表形式,但有的关系词语并非与一两种关系发生固定联系,不是典型标志,只能说是准标[1]。除关系标志这一核心构件外,标志构件还包括层次关系和复句类别两个配套构件,它们和关系标志一起完成对整个有标复句的管控。

层次关系的主要属性包括:a)层次数,标志一个有标复句由几层构成;b)单层关系,标志有标复句中某一层的关系;c)层关系词语,标志某一层的关系由哪组(个)关系词语来表达。有标复句句式类别由分句的第一层关系决定,第一层是什么关系就将整个有标复句标志为什么句式,因此,复句类别的属性值可以由层次关系的单层关系推导出来。

2.2　本体描述

本体捕获的下一步工作是本体描述,即用合适的描述语言来表达概念和术语[9]。在选择描述语言时,考虑到本课题的目的只是建立本体的静态元模型,并不需要描述动态交互,因而可以采用成熟度较高的 UML。当然,UML 是面向对象的建模语言,它与本体建模之间存在一种映射,对于具体的映射规则文献[11]作了细致的分析和说明,本文在此不作赘述,只是直接应用。结合文献[5,6]的成果和上文的句法分析,可以认定有标复句领域实义构件中需要描述的概念有五个,即词、短语、句子语气、分句和分句组;标志构件中需要描述的概念有三个,即关系标志、层次关系和复句类别。考虑到文章篇幅,本文只给出与文献[5,6]中描述有所不同的句子语气和标志构件中概念的 UML 类图及其说明,如图 1 ~4 所示。

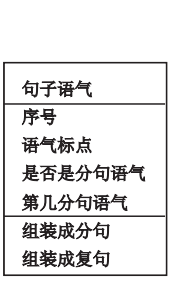

图1 句子语气类图

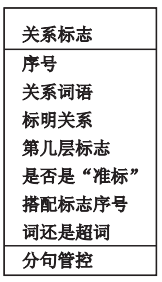

图2 关系标志类图

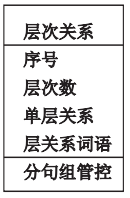

图3 层次关系类图

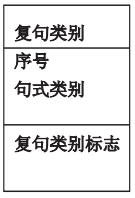

图4 句式类别类图

图 1 中,句子语气类的序号是指某个句子语气对象在复句中的线性位置,取值类型是整型;语气标点是指句子语气对象所使用的标点符号,取值类型是枚举型,取值范围 = {,、;、。、?、!};是否是分句语气是指句子语气对象是分句语气还是整个复句的语气,取值类型是逻辑型;第几分句语气是指当“是否是分句语气”取值为真,即是分句语气时,句子语气对象是第几分句的句子语气;句子语气类主要有两个操作,当“是否是分句语气”取值为真,即是分句语气时,“组装成复句”操作为空,当“是否是分句语气”取值为假,即是整个复句语气时,“组装成分句”操作为空。

图 2 中,关系标志类的序号是指某个关系标志对象在复句中的线性位置,取值类型是整型;关系词语是表示关系标志的词语,取值类型是字符串型;标明关系是指本对象所标明的分句之间的关系,取值类型是枚举型,取值范围 = {因果、推断、假设、条件、目的、并列、连贯、递进、选择、转折、让步、假转}[2];第几层标志是指关系标志在复句中所处的层次,取值类型是整型;是否是“准标”是指关系词语是否是典型标志,取值类型是逻辑型;搭配标志序号是指与此关系标志对象共同形成完整关系标志的标志对象的序号,取值类型是整型;词还是超词是指表达关系标志的是词还是比词大的超词单位,取值类型是逻辑型;关系标志的主要操作是分句管控。

图 3 中,层次关系类的序号表示其对象标志的是有标复句

中的第几个层次,取值类型是整型;层次数是指有某一标复句中共有几个关系层次,取值类型是整型;单层关系是指某一层次的关系类型,取值类型是枚举型,取值范围与关系标志中标明关系的取值范围一样。层次关系词语,表示某一层关系用什么词语来表达,取值类型是字符串型;层次关系的主要操作是分句组管控。

图4中,复句类别的序号实际上表示的是对象所属有标复句的序号,取值类型是整型;句式类别是指某一有标复句的关系类别,取值类型是枚举类型,取值范围与关系标志中标明关系的取值范围一样。复句类别的主要操作是复句类别标志。在考虑同类关系时,有两组泛化关系值得注意,如图5、6所示。

在泛化关系中,如果一般类特化出它的所有子类(不再有其他的子类)时,这种泛化称为完全(complete)泛化,如果存在某种具有公共父类的多重继承,这种泛化称为交叠(overlapping)泛化[12]。句子语气和复句语气、分句语气之间的关系就是完全泛化,如图5所示;而关系标志和词关系标志、短语关系标志、非短语超词标志之间就是一种交叠泛化,如图6所示。

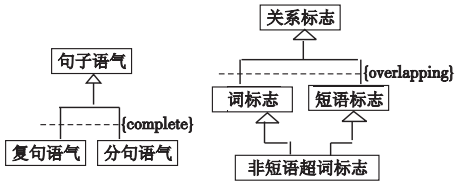

图5　句子语气类的完全泛化

图6　关系标志类的交叠泛化

考虑类间关系时,标志构件的类间关系是重点。如图7所示,在有标复句中,复句类别由层次关系中的第一层关系决定,所以复句类别与层次关系之间是依赖关系;而每层层次关系也最终由关系标志决定,因此层次关系与关系标志之间也是依赖关系;这种依赖关系的传递性也导致复句类别与关系标志之间是依赖关系。这也从侧面反映了关系标志在标志构件中的核心地位。

图7 关系标志、层次关系、复句类别三者之间的关系

2.3　本体集成

综合文献[5,6]的成果和本文的研究,可勾画出有标复句本体静态元模型,如图8所示。

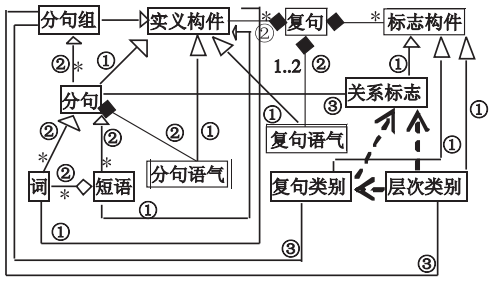

图8　有标复句本体静态元模型

图8中的"*"号表示多个意思,是"1…*"的缩写,如短语可由多个词组成,分句组可由多个分句组成等。为使图看上去更简洁,本文采用了缩写形式。图中①表示"is-a"关系,如分句是一种实义构件;②表示"part-of"关系,如实义构件是复句的组成部分;③表示"attribute-of"关系,如关系标志是分句的一个属性。

2.4　本体评估

在文献[9]中,Uschold 等人提出了一个对所构造本体进行评估的方法学框架,该框架列出了本体评估的三条基本原则:a)对本体应用领域的分类是否完整;b)本体描述与任务背景是否无关;c)新生本体和已有本体集成时重用率是否高。

依据上述三条原则对本课题构造的本体进行评估,可得出如下结论:a)以实义构件和标志构件来对有标复句构筑单位进行分类在语法学上是完整的,但如何将其映射到中文信息处理领域还需要进一步研究;b)本文用 UML 描述的有标复句本体的抽象程度是合适的,它既能保证与任务背景的无关性又可以为下一步的具体应用奠定基础;c)由于文献[5,6]和本课题的研究都是在相同的语法学框架内展开的,在本课题的研究中本体集成的重用率是较高的,但这也可能存在排斥其他理论框架内已有本体的风险。

3　结束语

有标复句是一个非常复杂的领域,对其中的概念进行分析并进行本体构造,目前在国内外都还只是处于初步阶段。本文在文献[5,6]的基础上尝试了这方面的工作,通过对标志构件的句法分析,在本体构造方法框架的指导下,用 UML 语言构造了有标复句领域相关概念的本体模型,并对它们进行了相应的评估。这些在缺乏任务背景的前提下构造的本体并不能马上应用于特定的任务,但它们的抽象程度以及和已有本体的兼容性是合适的,相信经过进一步修正和扩充,必将成为面向信息检索有标复句本体库的基础。

参考文献:

[1] 邢福义.汉语复句研究[M].北京:商务印书馆,2001.

[2] 姚双云.复句关系标记的搭配研究与相关解释[D].武汉:华中师范大学,2006.

[3] GUARINO N, MASOLO C, VETERE G. OntoSeek: content-based access to the Web[J]. IEEE Intelligent System, 1999, 14(3):70-80.

[4] SHUN S B, MOTTA E, DOMINGUE J. ScholOnto: an ontology-based digital library server for research documents and discourse[J]. Intl J Digital Libraries, 2000, 3(3):237-248.

[5] 胡金柱,王琳,肖明,等.汉语复句本体模型初探[J].华中师范大学学报:自然科学版,2005,39(4):466-469.

[6] 胡金柱,罗旋,肖明,等.本体论在复句领域概念建模中的应用[J].计算机应用研究,2006,23(10):212-214.

[7] GUARNO N, GIARETTA P. Ontologies and knowledge bases: towards a terminological clarification[M]//MARS N. Towards very large knowledge bases: knowledge building and knowledge sharing. Amsterdam:IOS Press, 1995:25-32.

[8] GRUBER T. Towards principles for the design of ontologies used for knowledge sharing[J]. International Journal of Human-Computer Studies, 1995, 43(6):907-928.

[9] USCHOLD M, GRUNINGER M. Ontologies: principles, methods and applications[J]. The Knowledge Engineering Review, 1996, 2(11):2.

[10] 邢福义.汉语语法学[M].长春:东北师范大学出版社,2000.

[11] 陈凯,何克清,李兵,等.面向对象的本体建模研究[J].计算机工程与应用,2005,20(2):40-43.

[12] 朱三元,钱乐秋,宿为民.软件工程技术概论[M].北京:科学出版社,2005.