汉语句子语义三维表示模型

朱 倩,程显毅,韩 飞
(江苏大学 计算机科学与通信工程学院,江苏 镇江 212013)

摘 要:如何表示和计算汉语句子的语义一直是自然语言理解的主要目标之一.在分析现有国内外关于语义表示研究成果基础上,提出了汉语句子语义的三维表示模型,即"义面—义原—义境"模型.该模型可以使句子包含的信息更准确、更全面地表示出来,为汉语语义知识建模和语义计算的研究提供一种新的思路.

关键词:自然语言理解;语义;义面;义原;义境
中图分类号: TP319.4 **文献标识码**: A **文章编号**: 1673-4785(2009)02-0122-09

A three-dimensional representative model of Chinese sentence semantics

ZHU Qian, CHENG Xian-yi, HAN Fei
(School of Computer Science and Telecommunications Engineering, Jiangsu University, Zhenjiang 212013, China)

Abstract: How to express and calculate the Chinese sentence semantic has always been one of the main goals in the aspect of natural language understanding. Based on the analysis of the research on the semantic at home and aboard, the three-dimensional model is proposed to express Chinese sentence semantic, that is "word semantics-word ontology-context" model. The model can express the information which were contained by the sentences more accurately. It is provide a new method for the research of making a model of Chinese semantic knowledge and calculating the semantic.

Keywords: natural language understanding; semantic; word semantic; word ontology; context

随着互联网以及大规模数据存储体系的迅猛发展,人类已经进入名副其实的海量信息时代.然而,拥有海量数据仅仅意味着人类拥有全面、深入、方便地驾驭这些海量数据中所蕴涵知识的潜在可能性.现实状况是,目前对海量数据的操作主要还在信息检索阶段,根本谈不上构建于其上的知识组织、总结及分析.彻底扭转此被动局面的惟一途径是,信息处理必须跨越到语义计算.

语义计算(语义分析)指的是将自然语言句子转化为反映这个句子意义(即句义)的某种形式化表示.即将人类能够理解的自然语言转化为计算机能够理解的形式语言,做到人与机器的互相沟通.然而,限于目前的技术水平,印欧语言在这方面的研究已经取得了一定的成果[1],但是对于中文的语义分析则相对落后.其中主要的原因在于,相对于印欧语言,中文没有那么丰富的形态变化,中文的词类与句法功能不是一一对应的,中文的词、短语、句子之间的界限是模糊的.除此而外,甚至可能是更重要的原因在于,中文的结构更加依赖于语义的制约[2].

1 传统语义计算模型

合适的语义表示是有效语义计算成功的一半.目前,有如下几种有影响的语义表示模型:

1)概念从属理论(conceptual dependency, CD).CD理论认为,人在理解自然语言时依赖的是潜在的概念表述,而不是具体的词或句子.人们总是用以前遇到的更简单、更基本的事来理解现在所遇到的事情.因此,当计算机理解自然语言时,也要依赖事件的概念表述而不是特定的词或句.概念

收稿日期: 2008-12-16.
基金项目:国家自然科学基金资助项目(60702056).
通信作者:朱 倩. E-mail: callmezq@126.com.

是指动作或在某一方向上对一物体做些什么,所有概念都可以由少量作用于物体的原语动作来描述,这是概念从属理论的基本思想[3]. CD理论希望对常识进行系统而又具体地描写,并利用原语来便利推理,从而达到对语言的自动理解.但从另一方面看, CD对常识描写是相当刻板和定式的.

2)语义场理论 (theory of field, TF). TF是介于单个词和整体词汇之间的一种活的现实.作为整体的一部分,它们与词一样具有被并入一个更大的系统中去的特征,而又和词汇一样,具有被分成较小单位的特性.

汉语语义分析的着眼点在于分析出句中所有概念之间的关系.关系语义场强调的正是义项之间的关系,因而对汉语的语义分析可以借鉴语义场理论[4].

3)格语法 (case grammar, CG). 格语法的基本思想是:动词在句中起中心作用,参与动作的各个体称为"语义格",且格的数量是有限的.针对每个动词的义项,由可能的"语义格"子集构成格框架,这一子集分为必要的和可选的 2个集合.

格语法最大的特点是承认语义在句法中的主导作用,由格语法分析可以得到句子的深层语义结构,给出各成分的语义角色,对于确定正确的句法结构有很大帮助.

格语法在汉语分析中存在以下 3个缺点:

①无法解决汉语的连动和兼语句式.格语法认为动词在句中起中心作用,那么分析句子时首先要确定句子的核心.汉语缺乏形态特征,作为核心的主动词通常也缺乏形态特征.如何在有多个动词的连动式和兼语式中找出句子的核心是汉语信息处理的一个难题,也是格语法无法解决的问题.

②短语内部各成分间关系无法确定.格语法提出的各种格关系都是名词性短语和动词之间的语义关系,对于名词性短语内部和动词短语内部各成分关系的确定没有给出.

③汉语词汇语义分类标准不确定.

4)知网 (hownet).知网是一个以汉语和英语词语所代表的概念为描述对象,以解释概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库.知网借鉴了概念从属理论的原语概念,提出了 1 500多个义原,用来描述概念、概念之间的关系及属性与属性之间的关系[5].

义原具有层次性,分为实体、事件、属性、属性值、数量、数量值、句法特征、次要特征和动态角色等类别.概念由义原描述,也具有层次性和分类.知网对每个事件义原给出了角色框架,列出了某一类事件发生时框架中的必要绝对角色.

知网适合于汉语的语义分析,但知网也有以下不完善的地方:

①知网强调了概念即义项之间的关系,但比义项更大的语义单位即语义块则没有提及.

②知网给出事物类概念和事件类概念之间的关系,但对事件类概念之间的关系没有给出,不利于解决汉语主动词的确定问题.

③知网提出了 73个语义角色来描述概念之间的关系,在事件义原的框架中以这 73个语义角色描述框架的必要语义角色;同时知网还提出了其符号系统,用以在词典的语义定义和事物类义原的共性描述中描述概念之间的关系.对概念之间关系的描述存在 2套描述体系,且这 2套描述符号并不一致,不能很好对应.

④概念之间的关系描述还很不够,需补充.

5)概念层次网络 (hierarchical network of concepts, HNC). HNC理论的目标是以概念联想脉络为主线,建立一种模拟大脑语言感知过程的自然语言表达模式和计算机理解处理模式,使计算机获得消解模糊的能力[6].

HNC建立的语句表述模式以句类为中心,所以可称之为句类体系. HNC定义的句类是指句子的语义类型,而没有陈述句、疑问句、祈使句和感叹句之分. HNC建立了句类的表示式,句类表示式由语义块构成,语义块是句子的下一级语义构成单位.语义块是句类的函数,也就是说,语义块的含义取决于句类,一个句子应该有几个什么样的语义块,这是由句类决定的. HNC发现,句类表示式存在有限的基元类型,总共有 57种,称为基本句类.

句类表示式是句子语义的基本框架,是句子所表达的基本语义信息.在语言理解处理中,判定句子所属的句类,并辨认出该句类表示式中的各个语义块,是句子理解处理的一项基本内容.因此,以 HNC的句类体系为指导来开展句子语义研究,具有十分重要的应用价值[7].

6)形式语义学 (逻辑语义学、自然语言逻辑).逻辑语义学着重研究自然语言这种符号系统的语义

模型. 20世纪 70年代初, 蒙太格 (R. Montague) 创立的蒙太格语法就是自然语言逻辑的开端[8].

形式语言的符号和它们所表示的概念之间的对应关系是确定的, 符号公式的结构反映它们的意义. 把一个理论形式化后, 就可以暂时完全撇开原来理论中的概念、命题的意义, 而只从语言符号、公式结构 (符号组合的形态) 方面研究. 意义是抽象的, 往往不容易精确地理解和掌握. 而符号和公式是有穷的具体的对象, 能够对其作更精确、更严格的研究, 从而通过对具体对象的研究把握抽象的东西.

以形式化为目标的语言逻辑的问世, 为人们提供了有效的参照系, 从而使人们对自然语言的复杂性有了比较清晰的认识. 但形式化方法在自然语言逻辑领域的缺陷与不足具体表现在以下 3个方面[9]:

①形式化在语言逻辑领域不具备普遍的效力;

②形式化方法不能彻底解决日常实际语用的恰当性问题;

③形式化方法不能取代自然语言的修辞现象.

尽管存在众多的语义计算模型, 研究人员也认识到了语境 (主体、上下文、常识、背景等) 在语义计算中的重要性, 但是目前还没有有效的结合语境的语义计算模型.

由于自然语言文本占据了互联网的大半河山, 同时, 在可预期的将来, 对声音、影像、图片的检索仍将严重依赖自然语言分析技术 (正如近两年 Google推出的图像与视像搜索引擎所做的那样), 语言计算的重要性也就不言而喻了. 可以预期, 它将成为信息科学技术中长期发展的战略制高点.

2 汉语语义

语言是人赖以从事复杂思维的工具, 思想是语言的内容 (语义), 两者相辅相成. 但语言既不等同于说出的话, 也不能等同于写下的句子.

语义研究是语言学研究的一个重要组成部分, 也是自然语言处理中不可忽视的研究内容. 语句所表达的意义分为句义和话语义 2部分.

2.1 句义

语言的一个主要用途是描述人的外部世界. 句子由字和词组成, 字和词都是音义结合体, 所以句子也就有了意义.

定义 1 句义是字义和词义根据一定规则组合的产物.

可以相对孤立地考察字义和词义, 比如在查阅字典时, 可以机械地研究组字成词, 组词成句后的句义, 而不必考虑句子使用时涉及的语境因素. 句义是一般的、稳定的意义, 浅层语义, 包括逻辑语义 (真值条件义) 和字面语义.

思想的、客观的、不包括人的主观因素的那部分内容, 被许多哲学家称为命题. 也就是说, 客观的思想以命题的形式出现, 一个命题或者真实地反映了外部世界的某个现象, 或者对某现象做出了不正确的、虚假的描述. 用逻辑的术语说, 前者为"真 (T)", 后者为"假 (F)", 真和假统称为真值, 所以, 命题具有真值. 命题虽有内容, 但无语音、语法外形. 所以, 从物理特性上看, 命题是与句子不同的概念, 因为一个命题可以由任何语言表达, 就是在一种语言里, 也可通过多种句法途径来表达同一个命题. 所以, 命题与表达命题的语言相对独立. 命题没有英语的还是汉语的区别, 因为它只有语义特征, 不具语言特征[8].

句子的真值条件是研究一个句子在什么场合下为真, 在什么场合下为假. 凡是陈述性的语句都具有真值条件. 如果把真值条件当作句子的逻辑语义, 就有了句子的真值条件义.

举例来说, 根据如下模型 $M = <D, F>$,

其中:

$D = \{$张三, 李四, 王五, 小兵, 大力, 陈规$\}$,

$F(a) =$张三, $F(b) =$李四, $F(c) =$王五, $F(d) =$小兵, $F(e) =$大力, $F(h) =$陈规,

$F(H) = \{$张三, 李四, 王五, 小兵$\}$,

$F(L) = \{ <$张三, 李四$>, <$王五, 小兵$>, <$大力, 陈规$>, <$张三, 王五$>, <$小兵, 大力$>, <$陈规, 大力$>\}$

因为 $F(a) =$张三 $\in F(H)$, 所以, $H(a) = T$; 因为 $F(e) =$大力 $\notin F(H)$, 所以, $H(a) = F$; 因为 $<F(e), F(h)> = <$大力, 陈规$> \in F(L)$, 所以, $L(e, h) = T$; 注意, 并不是在说, 陈述性句子的意义等同于真值条件义. 果真是那样的话, 必然会导致荒谬的结论, 即把所有取真值的句子视为同义, 把所有取假的句子也视为同义. 那么所以的陈述句总共就只有 2个句义了. 因此, 真值条件义只是陈述性句义的一个重要方面.

2.2 话语义

语言的另一个用途是表达使用者的情绪, 如:

1) 这音乐太棒了.

2) 但愿人长久,千里共婵娟.

它们不是对外部世界的描述,而是表达一种价值观或主观愿望,所以,这些语句无所谓真. 也无所谓假,没有真值条件.

语言的第 3个用途是运用语言办事,如:

3) 你被开除了.

4) 陪审团一致裁定:被告无罪.

类似"裁定"、"主持婚礼"、"结拜兄弟"等行为,必须在特定的场合下,遵循特定的方式,运用特定的言辞才能办到,离开了语言就办不成,被称为"言有所为". 一般来说,这种行为或是恰当的或者是不恰当的,也无所谓真假.

还有,句子的意思不能从字面得出,需要知识和推理才能决定,如:

5) 鸡不吃了.

它有 2个意思:"鸡吃饱了","不吃鸡肉了",理解"鸡不吃了"的语义要看上下文.

以上的 3种情形,统称为话语义.

定义 2　话语义是言者在特定语境中所表达的意义.

话语义可以等同于句义,但也可能超越句义,有额外附加的意义,甚至与句义完全不同. 话语义可被看成是由命题组成的集合,然而,只有说出的命题才构成话语义,此外却还有存储在记忆中未经表述的命题.

虽然话语义并不等同于承载该话语的句义,但听者总能从某个句义出发,借助知识,经推理而得到话语所表达的命题. 同时,它可以用与言者原来使用的句子不一定相同的一个或几个句子,把有关命题表达出来.

话语义是个别的意义、临时的意义,深层语义.

总之,句义虽不等同于话语所表达的命题,但命题总可以通过句子表达出来,否则便永远无法表达命题了[10].

3　汉语语义三维表示模型

3.1　模型描述

语言在本质上是主体以知觉的形式对世界的表现,它具体地反映了主体对世界的感受程度及其富有个性的呈现方式. 语言的排序深刻地揭示了主体对认识世界的认知模式. 也就说,完整地表达一个句子的语义信息,不但依赖于组成句子的词汇的义面信息(语言,词序),而且还包含用有限的义原信息(本体)表示每个词的深层信息(客体)及义境信息(主体、上下文、背景、常识等). 所以义面、义原、义境三位一体,不可分离. 句子语义的三维表示模型,类似于全息照片,可以使句子包含的信息更准确、更全面地表示出来,以便更精确地进行句子语义相似度计算(如图 1所示).

图 1中,X 轴表示词序列,是外延的(显性结构),称为语言空间、句子的表层结构;Y 轴表示词的义原信息,是内涵的(隐性结构),称为概念空间、句子的深层结构;Z 轴表示句子的背景信息,包括主体的思维状态和上下文,称为知识空间. X、Y 是客观的、静态的、语言的内部信息;Z 是主观的、动态的、语言的外部信息.

语言主要是用来交流思想的,在言语交际过程中,交际双方能直接凭感官得到义面,而最后得到的是句义和话语义. 义原和义境只是一个中介层面. 它们能帮助分词,得出句子的层次结构,以协助词义遵循义面提供的结构有规律地组合起来,从而得到句子的语义.

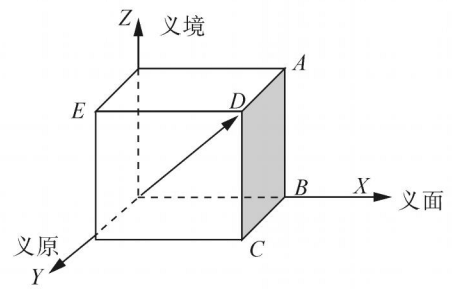

图 1　句子语义三维表示模型

Fig 1　The 3-dimensional representative model of Chinese sentence semantics

3.2　义面

在汉语语义自动分析中,词本身的语义信息是很重要的. 根据"组成性原则",句子的字面意义是由构成该句子的词的语义以及这些词之间的语义关系组成的. 因此,在语义分析中,义面(词汇语义)是基础,是一个汉语语言知识的词典,包括:词语、词性、词项、词例.

语言中的词汇具有高度系统化的结构,正是这种结构决定了词的意义和用法. 这种结构包括词和它的意义之间的关系以及个别词的内部结构. 对这种系统化的、与意义相关的结构的词汇研究叫做词

汇语义学. 从词汇语义学看来,词汇不是词的有限的列表,而是高度系统化的结构.

1)词语

选择词语的依据是建立大规模汉语语料库,按出现频率形成的词语表,而不是仅仅依据某一本现成的词典. 知识词典注意收集已经流行又有较固定可能的词语,如“因特网”、“欧元”、“下载”、“点击”、“黑客”等,但又不盲目求新.

2)词项

词项,在多数语言学文献称为义项,至所以用词项取代义项,是突出语言层面. 词项是对词语的定义,不同的词项给予不同的知识编号. 词项的选择要经过精心考虑. 一般很注意某一词项的现代的流通性. 例如“曹”在普通词典中至少有 2个词项,一是“姓”,另一是“辈”如用于“尔曹”. 而知识词典最好只选择第 1个词项.

3)词例

词例的选择主要是为那些具有多个词项提供例子. 这些例子的要求是:强调例子的区别能力而不是它们的释义能力. 它们的用途在于为消除歧义提供可靠的帮助. 这里试以“打”的 2个词项为例,一个词项是“买”,另一个是“辫编”,假设它们对应的知识编号分别为 1和 2,则对应知识 1,可以给出词例(符号“~”表示“打”字):~酱油,~张票,~饭,去~瓶酒,醋~来了;则对应知识 2,可以给出词例:~毛衣,~毛裤,~双毛袜子,~草鞋,~一条围巾,~麻绳,~条辫子.

4)词性

词性对句子结构分析和词项的选择有贡献. 好的词性标注就是通过采取适当的方法,根据上下文语境关系,消除句子中词的语法兼类,使得无论一个词兼有几种词性,在特定的场合下只保留其中最合适的一种.

3.3 义原

从语言理解的心理学出发,以人类共有的对某事物认识的概念出发,认为人们在认识客观事物过程中,存在着某种义原(本体、概念基),义原是最基本的、不易于再分割的意义的最小单位. 语言的理解过程就是把语句映射到义原的过程. 任何一种语言的词汇是离不开该种语言的,但概念是独立于语言的. 在这样一种思想知道下,词汇只是概念的符号,代表着一组可能知道的,用于该词汇所表达概念的所有特征.

从词汇引申到句子,句子被认为反映概念和概念之间的相互关系,并形成一种概念结构. 一个句子所包含的独立于语言的东西,不是句子的语法结构,也不是语义结构,而是它的概念结构. 2个句子只要含义相同,就有相同的概念结构. 所以,理解一个句子的关键在于提取句子的概念和概念结构.

以义原为基础的理解系统,词汇只是概念的符号,从整体而言,在这样的系统里,没有词只有概念. 最基本的概念集合组成义原集合,义原以动词为中心,相当于语言学研究者声称的语义角色.

语义角色是指有关语言成分在所指的语句所表达的事件中所扮演的参与者角色. 从某种意义上说,语义角色是语言学家对句子中有关结构成分之间的意义关系的一种分类. 这种分类的粗细程度,可以因语言学家的认识或具体的应用目标的不同而不同. 一般分 3个层次[11]:论元的语义角色(微观层次)、语义格(中观层次)、因—中心词—果(宏观层次).

1)微观层次

在这一层级上,又分为基于特定动词的角色和基于特定领域的角色 2种.

前者根据特定动词的意义所指,来确定其论元成分的语义角色. 比如“吃”的 2个论元的语义角色分别是:吃者,所吃. 这样,如果一种语言有 1 000多个动词,那么必将有 2 000多个语义角色.

后者根据各种特定的生活领域的具体场景,来确定有关场景要素的语义角色. 这种方法在目前的信息抽取中比较流行. 比如,在 Stallard[12]报道的机票订购信息系统中,有下列角色:出发城市、目的城市、起飞时间等.

2)中观层次

在 Fillmore[13]的语法理论中,先后用到下列格:

施事格(AGE):事件的发起者;

经验格(EXPER):经历精神或心理等事件的经验者;

受事格(OBJ):动作的承受者或状态;

源泉格(SOUR):物体移动的始点状态变化的初状态;

目标格(GOAL):物体移动的终点,状态变化的结果;

处所格(LOG):动作或状态发生的场所;

时间格(TME):动作和状态所进行的时间;

工具格 (NST):完成该动作所使用的工具.

很显然,这种层级的语义角色也是基于动词的.不过,它不是基于一个具体的动词,而是基于具有句法、语义共性的一类的动词.比如,表示运动的动词可能会涉及处所、源点和终点、或路径等语义格,表示转让的动词可能会涉及施事、与事和受事等语义格.这种语义角色可以说是语言学文献上讨论得最多的,但是,也是最难以给出合适的定义的语义范畴.以至于 Dowty[14]声称,"要想系统、一致地给所有动词的全部配项标明语义格,这在经验上几乎是不可能的".

3)宏观层次

鉴于中观层面上各种语义格的定义和区别的纠缠不清,清华大学孙茂松教授只对语料标注"因"和"果"2种语义角色;开发了一个 400万词规模的汉语语义骨架标注语料库,对每个句子标注有面向语义的因事 (S)、中心谓词 (V) 和果事 (O) 块信息.

义原除了指语义角色外还包含知识本体,有关知识本体的内容不在本文讨论.

3.4　义境

义境是语境的简称,语境有广义与狭义之分的说法.广义语境包括句子自身 (简称上下文)和句子形成过程的外部环境 (简称语域).对人类交际者来说,上下文与语域的区分是清晰的.上下文里蕴涵着现场语言信息 (简称言内信息),语域里蕴涵着语言之外的现场和相关的积累信息 (简称言外信息).言内信息与言外信息相互耦合形成交际语境,在此过程中,交际者得以实现对自然语言的理解[15].但是对当前的计算机来说,言外信息是不存在的,不具备言内信息与言外信息相互耦合的基本条件也就不可能形成交际语境;因此,必须在交际语境基础上进行一定的简化,从而形成计算机语境的框架形式,即交互语境的框架[16].具体的讲,义境研究如下问题:

1)常识

例:a　蓝球放在桌子上.

　　b　地球放在桌子上.

根据常识可以断定,句子 a可能为真,而句子 b一定为假.

人们至今不能确定,计算机究竟应该储存多少常识和专门知识,才能达到令人满意的自然语言理解水平.

2)上下文

指句子不是孤立的,需要推理才能确定其句义.自然语言的陈述中,也常常有很多不合语法、不合常理的地方,听者在解读这些陈述时会在生活常识的基础上自然地加以校正.

例:那天在商店里,看见一盆花,漂亮极了,但是价格很高,买回去恐怕要挨骂.

任何人听了都会在生活常识的基础上迅速地通过"联想、推理 (判断)与选择"理解为那天 (我)在商店里,(我)看见一盆花,(这盆花)漂亮极了,但是(这盆花的)价格很高,(如果我把它)买回去恐怕(我)要挨 (家里人的)骂.

没有生活常识和推理、判断能力的计算机能作出这样合乎情理的补充吗?实际生活中,需要"信息校正"的说法还在不断地产生.

3)主体的思维状态

例:张三相信李四喜欢王五.

　　张三想当老师.

　　他答应明天去办.

含有"信念"、"愿望"、"意图"、"规划"、"承诺"、"义务"等词语的句子涉及到主体的思维状态,也是一种语境.如何处理这样的语境也有一些研究成果,如:Church的内涵逻辑和分布式人工智能所研究的 Agent理论等,本文不做深入讨论.

4)领域背景

指说话者和听话者的知识水平要在同一个层次上,即"说清楚"是相对的.同样一种表述,对人是"说清楚"了,对计算机就往往"没有说清楚";对大人算是"说清楚"了,对小孩就往往"没有说清楚";对专家算是"说清楚"了,对一般人就往往"没有说清楚".一段正确的 C语言程序对于装备有 C编译器的计算机算是"说清楚"了,对没有装备有 C编译器的计算机就"没有说清楚".所以,"说清楚"的第一个必要条件就是"给出的信息所表达的内容能够和对方已有的知识相结合",能够为对方所"理解".

4　基于三维语义模型的汉语句子表示

4.1　语义表示的原则

汉语句义表示,主要遵循 2条最基本的原则,即组合原则和同构原则.

1)在语形方面,遵循结构主义的组合原则,根据组合原则,可将句子"我们遵循结构主义的组合原则"分层为"((我们)((遵循)((结构)(主义)

(的)((组合)(原则)))))))".

2)在语义方面,遵循蒙太古语法的句法语义同构原则(形式与意义同构):一个复合表达式的意义就是它的各部分意义的一个函项.对于外延性语言可以说,一个复合表达式的外延就是它的各部分外延的一个函项[17].

其核心就是认为复合表达式的意义只能由其组成部分的意义复合而成,所以语言研究的关键在于找出意义复合的方式,也就是所谓函数(即函项)式.

4.2 特定义境下的句子语义表示

虽然义面、义原和义境都对句义有贡献,但就目前技术水平而言,必须简化义面、义原和义境.义面暂不去涉及同指、量化、时态、语气、嵌套结构等高阶的相对深层次的语义现象;义原也只是讨论语义角色;义境只讨论上下文语境.所以,讨论建立在如下的假设下:在特定的义境下,义面和义原是惟一的.

所以在特定义境下有:

定义 3 句子的表层结构是指有序的线性结构,显示句子的前段信息、主题、后段信息内容之间的"主述关系".

句子的表层结构表示是义面表示形式之一,图2显示了1个句子的表层结构表示.

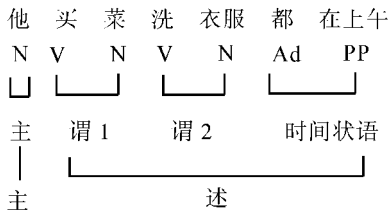

图2 句子的表层结构表示

Fig 2 Representation of sentence shallow semantic

定义 4 句子的深层结构是指显示句子轴心的谓语与周围体词短语之间的"句法语义"关系.

还以上句为例,图3显示了一个句子的深层结构表示.

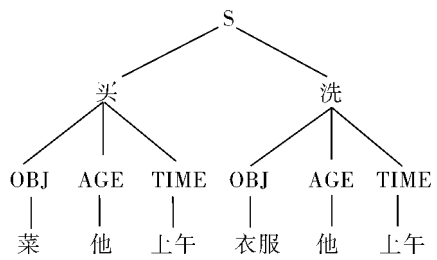

图3 句子的深层结构表示

Fig 3 Representation of sentence deep semanic

句子的深层结构表示也可表示为线性层次结构:

S((ACT(买)AGE(他)OBJ(菜)ACT(洗) AGE(他)OBJ(衣服))TME(上午))

表层的描述可以有多种,而深层的表示,理论上仅此一种.如上句还可以有下列多种不同的义面表示,但义原表示是同一个:

买菜洗衣服,他都在上午.

他买菜洗衣服,都在上午.

他买菜在上午,洗衣服也在上午.

他在上午买菜,也在上午洗衣服.

他都在上午买菜洗衣服.

所以,句子的深层结构表示可看作义原的一种表示.

4.3 三维模型的数据结构

把三维模型的数据结构定义为框架网,定义如下:

定义 5

框架网络::=框架*

框架::=<YM*(义面框架)><YY*(义原框架)>

YM::=<CY(词语曹)>&[<CX1*(词项曹)>]&[CL*(词例曹)]&<CX2*(词性曹)>

YY::=[BT*(知识本体曹)]&<JS*(语义角色曹)>

CY::=字符串(词的原型)

CX1::=字符串(对词语的定义)

CL::=字符串(词项的例子,为消除歧义提供可靠的帮助)

CX2::=N(名词)|V(动词)|Adj(形容词)|Ad(副词)|PP(介词)|P(代词)|

BT::=字符串(不可分割的最小概念集合中的元素)

JS::=字符串(语义关系表中的元素)

字符串::=词的序列(元语言符号)

"*"号表示可以重复,"|"表示"或","&"表示"与",[·]表示可选项,<·>为必选项,(·)表示解释内容.其中语义关系表为

连接关系:OR、NOT、SNCE、BUT、AND、MPLY、……

语义关系:AGT、OBJ、SOUR、GOAL、LOC、NST、TME、DAT、ASSOC……

形态关系:时态、句式、性、情态、时间……

属性关系:MOD、POSSESSIVE、HASPART、

SIZE、SHAPE……

量化关系:ALL、SOME、FEW、MANY……

语言符号:TOK(用于连接 YY和 YM)

集合关系:SUB、EQ、PART OF……

对于具体给定的系统,这些关系还可进一步扩充.

语义三维表示模型的数据结构是一个框架网.框架网由若干框架组成,每个框架视为一个义境,框架的曹即为义原,框架曹的值即为义面.

这种框架网很容易转换为面向对象的表示,为三维语义的实现提供了技术支持.

4.4　基于三维模型的语义分析

通过一个实例来说明基于语义三维表示模型的语义分析过程.

例如:张三 看见 了 李四 正在 酒吧 喝 酒

这个例子的简化框架网络表示如下:

YY1	TOK	YM1	YM1	CY	看见
	TM	PAST		CX2	V
	DAT	YM2	YM2	CY	张三
	OBJ	YY2		CX2	N
	LOC	YM6	YM3	CY	喝
	……			CX2	V
YY2	TOK	YM3	YM4	CY	李四
	TM	PROG PAST	CX2	N	
	AGT	YM4	YM5	CY	酒
	OBJ	YM5		CX2	N
	LOC	YM6	YM6	CY	酒吧
	……			CX2	N

有了这样的框架网结构,很容易写一个算法,形成词与关系的线性序列:

看见 DAT 张三 OBJ 喝 AGT 李四 OBJ 酒吧 LOC 酒

YY1　YY2　YY3　YY4　YY6　YY5

这就比源语言有了更明确的语义关系,实际上相当于理解为:“看见”有间接受事者“张三”,客体是“李四”,中心谓语是“喝”.施事者“李四”,具体地点“酒吧”,有关的“酒”.

4.5　基于三维模型的语义计算

表示的目的是为了计算,语义计算有 3层含义:同一个句子有几个意思(一句多义问题)?如何判断不同句子意思是相似的(2句相似问题)?如何判断真值相似而句义不相干(2句相关问题)?这 3个问题是句子语义计算的核心问题.

根据 4.2,在特定的义境下,可将句子映射为义原(深层结构).给出如下定义:

定义 6　设 Y_1、Y_2分别为句子 S_1和句子 S_2对应的义原.令 $\mu_{12} = Y_1 \cap Y_2$,如果 $\mu = 1$(完全重叠),则 S_1和 S_2同义;如果 $\mu_{12} = 0$(不交),则 S_1和 S_2不相干;其他情况的 μ_{12}即定义为 S_1和 S_2相似度.

定义 7　如果对句子 S_1存在 2个义境 J_1、J_2,它们分别对应句子 S_1的义原为 Y_1、Y_2,令 $v_{12} = Y_1 \cap Y_2$,如果 $\mu \neq 1$(不完全重叠),则 S_1一句多义.

有关 2个深层结构表示的交运算,将在后文讨论.

5　结束语

目前,我国的汉语研究还远远不能满足中文信息处理的需要.除了汉语书面语不分词连写、缺少形态变化、缺省部件等以外,汉语的语法研究一直受到西方语言学理论的影响,始终没能形成汉语自己的理论体系.总是带着印欧语的眼光来看待汉语、研究汉语,这就难免会削足适履.还有现有的面向中文信息处理的汉语研究主要不是面向语义,而是面向句法.这是现有汉语研究的一个最大缺憾,所以难以从真正意义上解决中文信息处理中遇到的问题.正因为现有的面向中文信息处理的汉语研究存在这样的问题,所以汉语语义的研究已经成为中文信息处理——一个阻碍信息社会经济发展的首要瓶颈问题.为此,汉语计算语义理论应该建立在内涵逻辑模型解释下的真正意义上的自然语言理解.也就是说,汉语计算语义理论研究主要不是着眼于句法,而是以语义分析为核心,辅以句法分析.要吸收句法分析研究的一切成果,但是面向中文信息处理的汉语研究的目标必须定位在语义分析,尽管目前的语义自动分析研究还困难重重,研究基础也极为薄弱.正所谓,取乎其上,得乎其中.必须把目标定得高些,当然,实际研究中,还必须脚踏实地,从基础做起.

本文基于假设“在特定的义境下,义面和义原是惟一的”讨论了语义三维模型的数据结构——框架网和单句的语义分析的基本原理.模型中所涉及的一些关键问题,如:义面到义原的映射问题、篇章的语义分析问题等将是今后研究的主要任务.

参考文献:

[1] JURAFSKY D, MARTIN J H. 冯志伟,孙　乐,译. 自然语言处理综论[M]. 北京:电子工业出版社, 2005: 80-92.

[2] KWONG O Y, TSOU B K. Semantic role tagging for Chinese at the lexical level[C]// Proceedings of IJCNLP 2005. Alicante, Spain, 2005: 411-416.

[3]由丽萍,范开泰,刘开瑛. 汉语语义分析模型研究述评[J]. 中文信息学报, 2005, 19(6): 57-64.
YOU Liping, FAN Kaitai, LIU Kaiying. Comment on models in Chinese semantic representation[J]. Journal of Chinese Information Processing, 2005, 19(6): 57-64.

[4]齐 璇, 马红妹, 陈火旺. 汉语的语义分析研究[J]. 计算机工程与科学, 2001, 23(3): 89-92.
QI Xuan, MA Hongmei, CHEN Huowang. Research of Chinese semantic analysis[J]. Computer Engineering & Science, 2001, 23(3): 89-92.

[5]董 强,郝长伶,董振东. 基于知网的中文信息结构抽取[EB/OL]. [2007-6-12]. http://www.keenage.com/html/c_index.html.

[6]林杏光. 用 HNC研究汉语为 NLP创立模式[J]. 汉语学习, 2002(3): 12-18.
LIN Xingguang. Application of HNC theory (a paradigm of NLP) to Chinese study[J]. Chinese Language Learning, 2002(3): 12-18.

[7]黄曾阳. 认知学与 HNC[EB/OL]. [2008-10-29]. http://www.hncnlp.com/.

[8]将 严,潘海华. 形式语义学引论[M]. 北京:中国社会科学出版社, 2005: 45-46.

[9]夏国军. 语言逻辑与形式化[J]. 南开学报:哲学社会科学版, 2004(3): 67-76.
XIA Guojun. The logic of language and formalization[J]. Nankai Journal: Philosophy, Literature and Social Science Edition, 2004(3): 67-76.

[10]缪建明,张 全. HNC语境框架及其语境歧义消解[J]. 计算机工程, 2007, 33(15): 10-13.
MIAO Jianming, ZHANG Quan. HNC context frame and context disambiguity[J]. Computer Engineering, 2007, 33(15): 10-13.

[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): 10-21.
YUAN Yulin. The fineness hierarchy of semantic roles and its application in NLP[J]. Journal of Chinese Information Processing, 2007, 21(4): 10-21.

[12]STALLARD D. Talk'n'travel: a conversational system for air travel planning[C]// Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP'00). Washington, DC, USA, 2000: 68-75.

[13]CHARLES F J, RUPPENHOFER J, BAKER C F. FrameNet and representing the link between semantic and syntactic relations[C]//Huang and Lenders eds. Computational Linguistic sand Beyond. Taibei: Institute of Linguistics, Academia Sinica, 2004: 231-238.

[14]DOWTY D. Thematic proto-roles and argument selection[J]. Language, 1991, 67(3): 547-619.

[15]黄曾阳. 语言概念空间的基本定理和数学物理表达式[M]. 北京:海洋出版社, 2004: 67-70.

[16]晋耀红. 基于语境框架的文本相似度计算[J]. 计算机工程与应用, 2004(16): 405-409.
JIN Yaohong. Text similarity computing based on context framework model[J]. Computer Engineering and Applications, 2004(16): 405-409.

[17]郭曙纶. 汉语计算语义理论及其原则[J]. 韶关学院学报:自然科学版, 2002, 23(6): 42-48.
GUO Shulun. The Chinese computational semantic theory and its principle[J]. Journal of Shaoguan University: Natural Science Edition, 2002, 23(6): 42-48.

作者简介:

朱 倩, 女, 1979年生,讲师,博士研究生. 主要研究方向为自然语言理解、数字图像处理.

程显毅,男, 1956年生,教授,博士生导师,博士. 主要研究方向为模式识别和自然语言理解. 发表学术论文 90余篇,其中被 EI收录 25篇,出版专著 2部.

韩 飞,男, 1976年生,副教授,硕士研究生导师,博士. 主要研究方向为智能计算、智能信息处理. 发表学术论文多篇,其中被 SCI收录 13篇.