文章编号:1000 - 582X(2004)04 - 0034 - 05

基于词联接的语义表示方法和知识库建设*

李良炎,何中市,易　勇
(重庆大学 计算机学院,重庆　400030)

摘　要:现有自然语言处理方法主要采取描述语言规律的基本思路,缺乏全面的语义表示能力,因此不能从语义层面有效处理各种类型的真实文本。笔者提出一种基于词联接实例的语义表示方法。该方法采取描述底层语言实例的基本思路,具有全面的语义表示能力,目前已应用于国家自然科学基金资助的"计算机辅助文学艺术创作研究——诗词曲联"项目。

关键词:自然语言处理;真实文本;语义表示;词联接实例
中图分类号:TP182　　**文献标识码**:A

自然语言处理方法实用化的前提是能够处理大规模真实文本。真实文本可以分为规范文本(如政府公文)和失范文本(如病句、习惯用语)、一般文本(如产品说明书)和修辞文本(如诗歌、小说)。修辞文本注重修辞手法的运用,具有很强的艺术感染力。规范文本和一般文本规律性强,处理起来比较容易,失范文本和修辞文本规律性差,处理起来非常困难。

当前自然语言处理的主要方法[1]在语言知识表示上主要采取描述语言规律的基本思路,语义表示能力不足,因此难以处理失范文本和修辞文本。符号处理系统方法运用规则描述语言规律,难以处理大规模文本中与已有规则相矛盾的语言现象。统计学方法运用概率描述语言规律,难以处理那些出现频率高但不合理的失范文本或出现频率低但合理的修辞文本。基于语法理论的方法不作语义表示,因此不能进行语义分析。基于语义理论的方法往往没有全面的语义表示,难以处理语义丰富的修辞文本。例如,情感语义是修辞文本具有强烈感染力的根本原因,而目前没有哪一种语义理论对情感语义进行知识表示。

实例方法是以具体的语言实例指导语言分析的方法,目前属于非主流自然语言处理方法。实例方法的优点是能够描述所有规则和不规则语言现象。不规则语言现象体现为语言实例,规则语言现象体现为语言规则,而语言规则可以通过语言实例归纳获得。很多研究者都逐渐认识到,实例方法是处理失范文本和修辞文本的有效方法。例如,语言学界有人提出将不规则语言现象放入"扩充的词库"[2];机器翻译研究中将实例方法与符号处理系统或统计学方法相结合从而提高效率[3]。与其它方法相比,实例方法更接近儿童的语言学习过程。儿童虽然只掌握了有限的语言实例和语言规则,却能参与日常语言交流。实例方法的缺点是语言实例的收集、筛选和管理工作量大。

笔者提出的基于词联接的实例方法采取描述底层语言实例的基本思路,具有全面的语义表示能力。语言实例是意义明确的真实文本,分为词联接、词组、短语、句子、段落、篇章等。孤立的词意义不明确,因此不是语言实例。词联接是两个意义明确且不相同的词按有序关系构成的语言单位,属于底层语言实例。其它复杂语言实例均由词联接构成,因此基于词联接的实例方法工作量最小。例如,现有实例方法一般以句子为单位[3],句子的数量远远大于词联接的数量。来自 N 个词的长度不大于 $k(2 \leqslant k \leqslant N)$ 的句子在理论上数量为

$$N^2 + N^3 + \cdots + N^k = \frac{N}{N-1}(N^k - N),$$

而来自 N 个词的词联接在理论上数量为 $P_N^2 = N^2 - N$。

* 收稿日期:2003 - 09 - 26
基金项目:国家自然科学基金项目(60173060)
作者简介:李良炎(1974 -),男,重庆开县人,重庆大学博士研究生,西南师范大学教育科学学院教师。主要研究方向:自然语言处理。

另外，该方法建立了完善的语义系统，提出了词的体验语义表示，从而为修辞文本的语义分析提供了重要的知识基础。

1　基于词联接的语义表示内容和方法

从信息论角度来看，语义是语言单位所蕴涵的信息。语言单位按层次由低到高分为字、词、语言实例（词联接、句子、篇章等）。低层次语言单位通过相互联系构成高层次语言单位。如果称语言单位的所有语义为固有语义，相同层次的语言单位相互联系产生的语义为上下文语义，语言单位整体具有的语义为整体语义，则语言单位的语义是一个系统（图 1）。由于字是最小语言单位，因此字的固有语义只包含字的整体语义。

语言单位总是具有一定的形式和内容。一般认为，语言的形式包括文字、字体、字号、发音等，语言的内容就是语言所指的事物或事件。然而这种看法忽略了修辞文本的一种特殊语义。如何解释精典文学作品对人的强烈感染力？只能认为一种被强化了的特殊语义在发挥作用。例如，“死亡、哭、美丽、高兴”等这类词具有明确的指称，同时具有强烈情感感染力。有必要表示出不同语言单位在这个方面的差异并让计算机能够识别。例如，日本开发的机器狗能够对人的不同语言产生不同情感体验并作出相应的动作反映。如果称语言的形式为形式语义，语言所指的事物或事件为指称语义，语言带给人的一般感受为体验语义，则语言单位的整体语义是一个系统（图 1）。字的整体语义中不包含指称语义和体验语义。

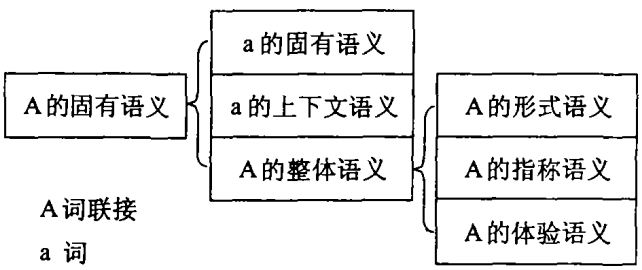

图 1　语言单位的语义系统（以由词构成的词联接为例）

现有自然语言处理方法的语义表示一般以词的固有语义为主要内容，忽略了其它更高层次的语义。不少学者发现了这一问题，并采取了补救措施：一是对词的固有语义进行扩展。例如，北京大学计算语言研究所开发了《现代汉语语法信息词典》[4]。词典中对词的前后照应和搭配等上下文语义进行了细致的刻划，已经被多家单位采用并取得了运用传统词典达不到的效果。二是引入短语或句子语义表示。例如，中国科学院声学研究所提出了抽象概念层次网络理论。该理论按语义对句子进行分类描述，作为词语搭配分析的依据[5]，目前在句子结构合理性识别和消解歧义方面取得了很好的效果。

根据语言单位的语义系统可知，句子的固有语义包括词联接的固有语义、词联接的上下文语义和句子的整体语义。在基于词联接的语义分析中，句子的整体语义和篇章的固有语义都是可计算的，因此基于词联接的语义表示主要内容为词联接的固有语义、词联接的上下文语义。

1.1　词联接的固有语义表示

词联接的固有语义包括词的固有语义、词的上下文语义和词联接的整体语义。词的固有语义是词在孤立状态下蕴涵的信息。由于字只具有形式语义，因此可以将字的语义放入词的形式语义，则词的固有语义只包括词的整体语义。词的整体语义分为词的形式语义、词的指称语义、词的体验语义。词的形式语义是词蕴涵的形式信息，主要包括词包含的字符、词的发音、词的构成方式等，在一般词典中有准确描述。

1.1.1　词的指称语义表示

词的指称语义是词蕴涵的指称信息。从逻辑学角度来看，词的指称语义是人从现实世界中认识到的具体概念和抽象概念。在人的认识中，现实世界的具体事物或事件是具体概念，具体概念通过概括化思维形成抽象概念，同一概念可以参与多个概括化过程。如果以概念为结点，概括化方向为边，则所有概念形成有向图（图 2）。如果进一步考虑概念之间全部关系（包括概括化关系）则所有概念构成非常复杂的网络。概念网络也就是语义理论中常常提到世界知识。目前的 HNC 理论[5]、英文 WordNet、中文 HowNet[6] 等研究都在建立概念网络。

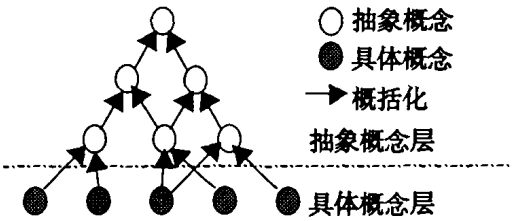

图 2　概念有向图

词的指称语义主要有 3 种表示方法：一是用词在现实世界中对应的事物作为词的指称语义[1]。这种方法直接面向具体概念，精度高但没有描述词与词之间的语义共性。二是用词的属概念作为词的指称语义，如词性分类表示[4]或词义分类表示。属概念是通过概括化获得的概念。这种方法描述了词与词之间的语义共性但没有有描述同一属概念下不同词之间的语

义个性,而且同一词有多个属概念时难以选择。三是词汇分解学派方法,用一组有限的语义基元通过组合操作来表示词的指称语义[5-6]。语义基元是构成复杂语义的元素,其组合可以描述非常复杂的语义。但现实世界是复杂的,要从中提取语义基元并得到大家公认是很难的。

传统词典中对词的指称语义表示主要采用内涵定义和外延列举的义项解释方法。从逻辑学来看,内涵定义分为词的属概念和词的种差。词的属概念就是词指称的具体概念的概括化,属于抽象概念。词的种差就是词与同一属概念下其它词的区别,可以是抽象概念或具体概念。为了避免第一种语义表示方法需要用到具体概念的缺点,可以在内涵定义中用种差的属概念来代替种差,从而形成一种属概念加抽象描述的内涵定义方法。这种方法在只运用抽象概念的条件下尽可能准确地表示了词的指称语义。

由于人认识的对象都来自现实世界,因此不同人之间具有公共概念子图。由于不同人对现实世界的认识深度和广度不同,概括化标准和方向也不尽相同,因此每个人都具有私有概念子图。人的语言交流主要建立在公共概念子图基础上。如果能够找到所有人在抽象概念层(图2)中的公共抽象概念子图,就可以用于属概念加抽象描述的内涵定义方法。由于同一概念的不同概括化方向可以分为两个不同的概念(因为其属概念不同),因此公共抽象概念子图可以转化为公共抽象概念子树。词的指称语义表示的形式化系统就是公共抽象概念子树,简称抽象概念树。抽象概念树的建立是一个非常复杂的知识工程,关键在于概念划分标准要能够被大数人认可。其基本方法是:

1) 确定位于抽象概念树根结点的抽象概念。该概念是一切事物的抽象,是一切概念和自身的属概念。例如,确定该概念的名称为"现象"。

2) 按通用标准对位于根结点的抽象概念进行划分。如"现象"可以分为"事物"和"特征"两个概念。

3) 按通用标准对子概念进一步划分,直到一定的层次为止。理论上概念划分可以无限地进行下去,但划分并不是越细越好。因为从心理学角度来看,人对事物的认识精度实际上是有限的,只要满足应用需要即可。目前已经建立的抽象概念树共有275个抽象概念,最多有11层。

1.1.2 词的体验语义表示

词的体验语义是词蕴涵的人对事物的体验信息。体验的本质是人对事物的感受,表现为人的情感和态度。事物是词的指称,由于词激发了人对该事物的联想,从而激发了人对该事物的体验,这是文学作品能够感动读者的基本机制。因此人对事物的体验可以看作人对词的体验。尽管不同人对相同的事物可能产生不同的体验,但作者用能够感动自己的语言去感动大多数读者,这一事实说明不同的人对相同词的体验基本相同或相近。因此词的体验语义表示以大多数人的体验为标准。

词的体验语义分为意味语义和情感语义。词的意味语义是词所指事物的形式因素如数量、形体、色彩、声音、味道、重量、力量、节奏、韵律、速度、质感等带给人的体验。从艺术美学角度可以看作事物的形式美感,一般分为优美和壮美。具有数量少、形体小、色彩浅、声音柔、味道淡、重量轻、速度慢等形式特征的事物更容易引起人的优美感受,反之则更容易引起人的壮美感受。词的情感语义是词所指事物的价值因素如道德、审美、功利等带给人的体验。人的情感非常复杂,主要采用语义基元组合的表示方法,用情感基元组合描述各种丰富的情感。

情感也可以看作概念,按一定标准对情感进行划分可以获得情感基元。人的情感总是由于人一定的自身状态、人与环境的利害得失、人对事物性质的评价和态度相联系。以这些因素为标准划分出的情感基元有:

1) 振、沉:描述人的情感基调。积极情感为振,消极情感为沉。

2) 闲、寂:描述人在无外界因素影响下的情感。肯定情感为闲,否定情感为寂。

3) 乐、哀、幸、惜:描述人由于利害得失而产生的情感。人有所得则乐,人有所失则哀;人该失而未失则幸,人该得而未得则惜。

4) 恋、伤、盼、忧:描述人由于过去和将来事物而产生的情感。人肯定过去则恋,人否定过去则伤;人肯定将来则盼,人否定将来则忧。

5) 爱、憎、怜、惧:描述人由于对象属性而产生的情感。人肯定对象属性则爱,人否定对象属性则憎;对象属性弱于人则怜,对象属性强于人则惧。

词的体验语义形式化对于修辞文本处理十分重要。因为文学作品中运用大量的修辞手段目的就是为了使人们欣赏作品形式、情感被作品内容所激发。计算机如果能够通过词的体验表示知识掌握人的基本体验规律,也就能够在一定水平上理解文本的体验。例如,在计算机诗词理解中要对诗词的风格和情感分析必须依赖于这些知识。

1.1.3 词的上下文语义表示

词的上下文语义是词在词联接中的相互关系，不能离开词联接而存在。主要包括词的角色语义、词的功能语义。词的角色语义是词在词联接中的语义地位，包括作用、施事、受事、工具等，主要用于句子合理性分析。词的功能语义是一个词对另一个词的功能语义和体验语义的影响作用，分为强化、无关和弱化，是句子语义整合计算的重要参数。例如，“非常”强化“好”、“不”弱化“好”。词的形式语义和指称语义是客观的，因此不受影响。

1.1.4 词联接整体语义表示

词联接整体语义是词联接作为一个整体蕴涵的信息，不能离开词联接而存在。主要包括位置语义、关系语义、合理性语义、形式语义、指称语义、体验语义。词联接的形式语义、指称语义、体验语义通过词的语义计算获得，因此不需要表示。词联接的位置语义是词联接的两个词在句子中的位置关系，主要通过语料学习获得。

词联接关系语义即两个词之间的关系，分为构成关系和语义关系。词与词的构成关系是词在句子中的地位关系，分为3种：偏正关系（很－好）、正偏关系（看－书）、并列关系（花－草）。词与词的语义关系是词的指称语义之间的关系，分为8种：修饰关系（性质与事物，如善良－人）、说明关系（条件与事物，如汽车－运输）、种属关系（具体与抽象，如男人－人）、领有关系（部分与整体，如手－人）、相等关系（同一事物，如中国－中华人民共和国）、相似关系（性质相近，如美丽－漂亮）、相反关系（互补关系，如对－错）、相对关系（互斥关系，如香－臭）。

词联接合理性语义即词联接在语言实际中被认可的范围，主要分为常识范围、专业范围、个人范围。词联接只有在特定范围中被承认才是合理的。依据是词联接的语义合理性或语用合理性。真实文本中大多数词联接都具有语义和语用双重合理性。修辞文本中有很多词联接只在个人领域中有合理性。

1.2 词联接的上下文语义表示

词联接的上下文语义是句子中词联接的相互关系。两个词联接通过公共词发生联系，进一步构成句子。因此任何句子都可以看作以词为结点，以词联接为边的图。词联接的上下文语义表示就是确定句子的图结构。对所有句子都进行词联接的上下文语义表示是不现实和不必要的。只能选择合适的语料，对语料中的句子进行表示，然后通过机器学习获得词联接的上下文语义知识。在语义分析中可以运用这些知识分析句子的词联接上下文语义。

2 基于词联接的语言知识库建设

任何自然语言处理的知识工程都是费时费力的，基于词联接的语义表示也不例外。语言知识的获取历来有理性主义和经验主义之争，可以靠语言学家的直觉或语感来获取，也可以从语料库大量可观察的语言事实中归纳出来[7]。基于词联接的语言知识库建设综合两种方法，初期主要依赖专家知识，随着实例知识和规则知识的积累，可以采取机器学习减少人的工作量。以下主要介绍基于词联接的语言知识库主要结构和知识获取方法。

2.1 字典（汉语特有）

字典存储字的固有语义，如字的字形、异体字、拼音、声调、音韵等。权威专家编撰的字典是获取知识的主要途径。

2.2 词典

词典存储词的固有语义。单义词具有一项固有语义。多义词则有多项固有语义。词的固有语义以指称语义为核心，不同指称语义可能与不同的发音和体验语义相对应。词典知识的获取来自词联接实例的学习。

2.3 词联接库

词联接库存储词联接的固有语义。相对词典来说，词联接库为语言分析提供了更丰富更明确的语言知识，这是基于词联接的语言知识表示以词联接库为重心而不是以词典为重心的根本原因。词联接库知识的获取主要来自受标语料学习和专家管理。自动提取和分析受标语料中包含的词联接固有语义是主要方式。专家定期对词联接表中的语义知识进行审阅、补充和修改是重要的辅助方式。

2.4 受标语料库

受标语料库存储大量受标语料。受标语料是经过语言专家理解并按照一定的语言知识表示内容和方法进行标注的真实文本。受标语料库是语言知识获取的重要来源。为了获得足够的语言知识，受标语料库应具有相当大的规模，这给语言专家带来很大的工作量。

人机结合的方式进行语料标注是语料库建设的一种好方法。基本过程是：首先选择一批语料，由专家独立标注完成，然后由计算机学习并获得语言知识。在标注第二批语料时先由计算机运用已有的语言知识对语料进行自动标注，再由专家检查、校正、标注。这种方法避免了语言专家的重复劳动但必须结合语言分析算法。

2.5 规则知识库

规则知识库存储语言规则。基于词联接的自然语言处理方法属于实例方法,但并不排除符号处理系统的规则方法。一方面可以从语言实例中提取语言规则,另一方面语言专家掌握大量合理而且容易运用的语言规则。将这些规则直接存储在规则知识库中,在语言分析时能够提高处理效率。例如,用于检验诗词作品形式和语义合法性的谱式知识就属于语言规则。规则知识的获取主要靠语言专家制定和计算机从语言实例中提取。

3 结 论

基于词联接实例的语义表示方法以词联接为语义表示的基本单位,具有全面的语义形式化系统。该方法首次将语言的语义分为形式语义、指称语义和体验语义,并对语言的意味和情感等体验语义进行了形式化,从而为修辞文本处理提供了重要的语义知识基础。中国传统的诗词文学语言是典型的修辞文本,诗词语言处理向传统自然语言处理方法提出了严峻挑战,同时也带来反思和发展的契机。基于词联接实例的语义表示方法形成于"计算机辅助文学艺术研究——诗词曲联"项目,目前已应用于该项目的诗词知识库建设中。

参考文献:

[1] 翁富良,王野翊. 计算语言学导论[M]. 北京:中国社会科学出版社,1998.

[2] 徐杰. 普遍语法原则与汉语语法现象[M]. 北京:北京大学出版社,2001.

[3] 冯志伟. 计算语言学基础[M]. 北京:商务印书馆,2001.

[4] 俞士汶, 朱学锋, 王惠,等. 现代汉语语法信息词典详解[M]. 北京:清华大学出版社,1998.

[5] 黄曾阳. HNC(概念层次网络)理论[M]. 北京:清华大学出版社,1998.

[6] 董振东,董强. 知网简介[EB/OL]. http://www.keenage.com/html/c_index.html, 2002.

[7] 黄昌宁. 统计语言模型能做什么? [J]. 语言文字应用, 2002,(1):77-84.

Semantic Representation Approaches and Corpus Construction Based on Term Connections

LI Liang-yan, HE Zhong-shi, YI Yong

(College of Computer Science and Technology, Chongqing University, Chongqing 400030, China)

Abstract: The current Natural Language Processing (NLP) means is chiefly grounded on the principle of language rules' description pattern, which lacks comprehensive semantic representation capability, and therefore it can not efficiently process the real texts in the layer of semantic meaning. This paper presents a semantic representation approach based on term connection samples. This method promotes the idea of describing fundamental language samples, and can analyze comprehensive semantic representations. So far it has been applied in project CAPC (Computer Aided Poetry Composing) funded by the Chinese Natural Science Foundation.

Key words: NLP; real text; semantic representation; term connection samples

(编辑 吕赛英)