一种基于修饰关系的自然语言语义分析方法

田卫新[1,2] 朱福喜[1] 但志平[2]
(武汉大学计算机学院 武汉 430072)[1] (三峡大学电气信息学院 宜昌 443000)[2]

摘 要 自然语言语义分析是自然语言处理技术走向深层应用的瓶颈。当前在概念、关系层次上的语义分析方法主要有两种:基于统计的特征向量抽取方法和基于语义词典(WordNet、HowNet等)的语义相似度计算方法。对于具体应用这两种方法都具有较大不足,前者由于统计模型的关系只适用于段落、篇章或多文档等粗粒度的语义分析,而不适合在句子词汇一级的应用;后者能方便处理实体概念之间的各种关系,但是如果想正确处理真实文本中的复杂修饰关系如概念与事件、概念与概念修饰、事件与事件修饰等关系,还需对语义词典和计算方法做进一步的扩展。提出了按照真实文本语句中词语之间修饰关系建立知识库,并设计了根据该知识库中已有修饰关系计算未知关系的算法;提出了可以依照修饰关系建立自然语言构句法的思路并给出了相关算法;最后给出了在语义分析系统上的实验,结果证明该方法是有效的。

关键词 自然语言处理,语义分析,修饰关系,知识库
中图法分类号 TP18 **文献标识码** A

New Approach to Analyzing Meaning of Natural Language on Modifying Relations

TIAN Wei-xin[1,2] ZHU Fu-xi[1] DAN Zhi-ping[2]
(School of Computer Science, Wuhan University, Wuhan 430072, China)[1]
(College of Electrical Engineering and Information Technology, Three Gorges University, Yichang 443000, China)[2]

Abstract Acquiring the meaning of natural language is a bottleneck to make deeper use of natural language processing (NLP). There are two main measures on analyzing meaning of natural language at conception-relation level: one is the method of extracting characteristic vectors based on statistics, and the other one is method of computing semantic similarities according to semantic dictionary like WordNet or HowNet. Both of the two methods have weakness when putting them to applications. The previous is only applicable to analyze the meaning of those materials with big granularities such as paragraphs, documents or multi-documents, but is not fit for the applications at the level of sentences or words. The latter can deal all sorts of relations between conceptions easily, but when coming to complicated modified relations between conceptions and events, conceptions and conceptions or events and events, the semantic dictionary and computing method shall be extended. This paper presented a new method to structure semantic knowledge base(SKB) according to the modifying relation of real context; algorithm of computing unknown relations on the knowledge base was presented; we pointed out the way to design the rules of constructing natural language sentences under modifying relation and present the algorithm; in the end we made experiment on the platform developed in the light of the theory mentioned above and the result shows the theory is feasible.

Keywords NLP, Syntax, Semantic, Modifying relations, Knowledge base

自然语言是人类知识的重要载体。随着计算机和互联网技术的不断普及,高可用的自然语言电子文本信息变得日益庞大,对这些信息的有效分析利用已经成为信息社会发展的主要制约因素。然而,一方面,面对海量的电子文本信息,精确检索、内容摘要、观点分析、主题分类等常规知识发现技术越来越依赖于对自然语言的深入处理;另一方面,由于自然语言表意的灵活性和复杂性,现有的理论不能充分利用蕴含在自然语言文本中的大量相关知识,造成了在自然语言处理领域常常面临知识缺乏的问题,而使得建立在其上的应用效果不能令人满意。

人类很早就认识到事物之间存在比较直观的层次关系,如:猫→动物,桌子→家具等。基于这种直观的层次关系,研究者们设计实现了语言知识库,比较著名的有 WordNet, HowNet 等。WordNet[1]是普林斯顿大学 G. A. Miller 等人设计并实现的英语词汇知识库。该知识库由节点和连接关系组成。节点是由同义词集(Synset)构成,同义词集内的元素是相互具有可替换性的词(Word)。由于存在一词多义现象,因此同一词可能会出现在多个同义词集中。词是知识库的最

到稿日期:2009-07-07 返修日期:2009-09-15
田卫新(1974—),男,博士,讲师,主要研究方向为机器学习、自然语言处理、文本挖掘,E-mail:t_wxin@hotmail.com;朱福喜(1957—),男,博士,教授,主要研究方向为人工智能、数据挖掘;但志平(1976—),男,讲师,主要研究方向为算法设计。

小单位。WordNet 最初版本的类型只有名词，在后来版本中逐渐加入动词、形容词、副词等实词类型。节点之间的连接关系主要有同义关系、反义关系、上下位关系、部分关系等，在不同的实词类型之间不存在连接关系。WordNet 的构建主要是手工完成的，可以利用文本中的一些语言特征来设计算法以提取某些关系[2,3]，算法最终需要经过人工筛选后才能存入知识库。该知识库主要用在信息检索、自然语言理解等方面，相关应用可参考文献[4-7]。HowNet[16]是由董振东等人设计的基于汉语的语义知识库。与 WordNet 不同的是，在该知识库中词使用更小单位的义原来表示。义原在 HowNet 中是不可再分的最小语义单位，根据构建者对客观世界的认识和理解预先定义。词或者概念使用义原和角色关系来描述。HowNet 概念之间的关系有上下位、同义、反义、部件-整体、施事-事件、受事-事件、事件-角色等 16 种。该知识库同样也是由手工构建的。与 WordNet 相比，HowNet 中的关系划分得更详细，因此在进行语义处理时相应能达到更好的效果。当前基于 HowNet 上的应用主要集中在信息检索、句法分析、数据挖掘等领域，相关研究可见参考文献[17-20]。

FrameNet 是由 Fillmore 等设计的一种建立在框架语义学理论基础上的知识库工程[8]。FrameNet 的目的是将英语中的每个词的每一个意义在真实语境中的语义和句法之间可能的搭配关系(Valence)存放起来。FrameNet 数据库中主要存放词条(Lexicon)、框架(Frame Database)和经过标注的例句(Annotated Example Sentences)。框架语义学理论认为词的意义在具体语境(框架)中才能得到体现。在精心筛选定义框架内使用计算机形式化分析处理自然语言语义是可行的，工程的主要难度在于规模问题。目前 FrameNet 已经定义了超过 10000 个词汇单元(Lexical Unit)，并对其中 6100 多个进行了完全标注，筛选定义了 825 个以上的语义框架(Semantic Frame)和 135000 个经过标注的例句。目前已经有很多基于 FrameNet 的研究报道，主要集中在自然语言理解、机器翻译等领域，相关研究见参考文献[9-12]。

除了上面提及的语言方面的知识库外，关于知识库方面的建设和应用，相当多的研究工作集中在某一领域的知识系统建设和推理上面，如领域本体的自动获取等，相关的报道见文献[13,15]。

根据对上面各种知识库的研究，本文提出了按照真实文本中出现的词之间的修饰关系建立语言知识库的方法。这种方法基于这样的认知假设：人类通过概念的内在属性和外部表现来认识概念以及多概念之间的相互关系。概念的内部属性和外部表现可以通过自然语言文本词之间的修饰关系获得。依照修饰关系可能产生的不同的语义效果，定义了 16 种修饰关系。另一方面，人对自然语言的理解是建立在一定的构句知识和对概念的认识基础上的。为了让计算机模仿人类在已有知识(修饰关系)的基础上获取新的知识(修饰关系)，对汉语的构句法做了一定的研究。下面对计算语言学语法方面的文献进行简要综述。

适合计算机处理的语言学语法理论最早是由美国语言学家乔姆斯基提出的短语结构语法(Phrase Structure Grammar)[14]。短语结构语法并不限于任何一门具体的语言，开创了形式化研究自然语言的先河。短语结构语法的目的是要以公式化的方法把一门语言中符合语法和不符合语法的句子区分开。语法将语言看成是按照某种生成规则形成的字符串，认为在合适的语言处理层次能找到句子的生成规则，从而按照该规则生成该语言符合语法的句子[21]。在自然语言处理上，一般在按照词性区分的句子成分级的层次上应用该语法。经过多年的实践研究，现在一般认为能适用一门完整自然语言的生成规则是不存在的，该语法易于计算机处理，实用性强，因而在工程上得到了较广泛的应用。针对短语结构语法分析效率较低的问题，美国语言学家布南雷斯(J. Bresnan)和卡普兰(R. M. Kaplan)提出了词汇功能语法(Lexical Functional Grammar)；针对短语结构语法存在的处理层次单一等缺点，美国计算语言学家马丁·卡依(Martin. Kay)提出了功能合一语法(Functional Unification Grammar)。此外，1985 年盖兹达(Gerald Gazdar)等人提出了广义短语结构语法(Generalized Phrase Structure Grammar)，1987 年普拉德(Pollard)等提出的中心词驱动短语结构语法(Head-Driven Phrase Structure Grammar)等体现了在语法分析基础上对语义的重视[22-25]。

上述各种语法理论本质上都是基于语言结构规则的。本文提出了一种基于语义驱动的句子结构分析方法，具体包括：1)将句子中存在修饰关系总结为 16 种类型；2)按照词间修饰关系建立知识库并用来分析句子语义；3)建立了汉语中基于词间修饰关系的构句法；4)在构句法的基础上，设计了根据已有修饰关系获取句中修饰关系、根据修饰关系生成句子的算法。

本文第 1 节介绍词间的修饰关系；第 2 节介绍基于修饰关系的句子语义分析；第 3 节介绍语句分析算法；第 4 节说明测试过程及结果分析；最后是结论和今后的工作展望。

1 词间的修饰关系

1.1 修饰关系概念

词之间的修饰关系是指在有明确语义的句子中一个词对另外一个词的描述、限制或支配等关系。例：

a. 海港的灯火闪着微红的光影。

句中(海港→灯火)、(闪→灯火)、(微红→光影)、(光影→闪)等即为修饰关系。

这种修饰关系不仅是一种语言现象，同时也深刻地反映了客观世界概念之间的各种关系。而从某种意义上说，概念及之间的相互关系构成了知识。看下面的例子：

b. 学校 开展了向英模学习的主题活动。

c. 中国石油 开展节能宣传周活动。

例子中“学校”、“中国石油”都和“开展”这一动作有关系，表明了在客观世界中“学校”、“中国石油”这两个实体都具有“开展”这样一种动作的能力。相反，当在客观世界中两个概念不存在某种关系时，我们在有意义的真实文本中也不会找到相应的修饰关系。如：

d. 一株水仙花 吃了一头大象。

e. 冰冷的 篝火映在每个人的脸上。

这样的句子，因为“水仙花”不具备“吃”的能力；“冰冷”也不是“篝火”应有的属性，所以在真实文本中是很难找到的。词之间的修饰关系和概念之间的相互关系的这种对应，让我们可以将对概念之间相互关系的处理转换成为词之间修饰关系的处理。

当某两个概念具有相同关系时，我们认为这两个概念是相似的，如“学校”和“中国石油”都和“开展”有相同关系，若它们还都和“举办”、“组织”、“美丽”等有相同的关系，那就认为“学校”和“中国石油”相似度较高，如果两个概念相联系的所有关系全都相同，则认为这两概念是等价概念。但是，“学校”和“中国石油”肯定会有不同的表现。正是基于这样的思想，我们设计了相关的算法，将在第 2 节详细介绍。

并不是所有的概念之间的关系都可以和词之间的修饰关系对应起来。可以认为词之间的修饰关系是概念之间关系的一个子集，因为人对客观世界的认识是由简单到复杂，由具体到抽象的过程。认识过程中形成的概念以及概念之间的相互关系因此存在类似的层次。两词之间的修饰关系通常只反映概念之间最简单、具体的关系，对于复杂关系或抽象关系往往需要多个词组成句子来表示。如下例：

f. 蓝藻是一种单细胞体植物。

g. 计算机系统是由软件和硬件构成的。

两个句子分别表达了两种抽象关系：“蓝藻”和“植物”之间的类属关系，“计算机系统”和“软件”、“硬件”之间的组成关系。这类抽象关系在语句中没有修饰关系直接对应，是对整个句子语义分析的结果。

抽象关系具有很强的概括性，对这种关系的获取是一种集中体现人类智能的活动。这种关系的获取是以足够的较简单、具体的关系为基础的。如上面 e，f 句必须是在具有丰富的关于植物和计算机系统相关知识的前提下才有可能得出这两种抽象关系，通常是由植物学和计算机科学领域的专家总结。以 f 句为例，要得到这样一种类属关系，首先必须对蓝藻的内在结构、外观、特征、表现行为等有全面的认识，然后将这些属性分别和生物、动物、植物的属性进行比较，才能最终得出结论。而只要全面认识事物联系的简单、具体的关系，就能够得到抽象的关系。随着认识加深，这些抽象的关系可能是另外更加抽象关系的基础。

因此从理论上讲，以词之间的修饰关系构建的知识库可以覆盖所有的概念及其关系，简单、具体关系可以在知识库中找到直接对应，复杂、抽象的关系可以在知识库基础上通过推导得出。但在实际应用中很难获得范围全面的知识，所以在确定句子的深层语义时应直接使用已有的抽象关系。

定义 1（词间的修饰关系） 即出现在真实语句中两词之间的一种偏序关系，该关系反映了其中一词对另一词的描述、限制、说明、支配、结构等语义现象。

1.2 修饰关系的类型

根据真实文本中出现的实词之间的各种关系，并考虑表意虚词在句中的结构支配作用，将词间的修饰关系概括为以下 16 种关系类型。

a. 动作（主体词和动作词之间的关系，修饰方向由动作词指向主体词）；

b. 变化（主体词和变化词之间的关系，修饰方向由变化词指向主体词）；

c. 述态（主体词和述态词之间的关系，修饰方向由述态词指向主体词）；

d. 情态（述态词和情态词之间的关系，修饰方向由情态词指向述态词）；

e. 修饰（目标词和修饰词之间的关系，目标词可为动、名、形容词等词性，修饰方向由修饰词指向目标词）；

f. 限定（目标词和限定词之间的关系，目标词可为动、名、形容词等词性，修饰方向由限定词指向目标词）；

g. 度量（目标词和度量词之间的关系，目标词为名词性，修饰方向由度量词指向目标词）；

h. 数目（数目词和度量词之间的关系，修饰方向由数目词指向目标词）；

i. 时间（时间词和动作词之间的关系，修饰方向由时间词指向动作词）；

j. 空间（空间词和动作词之间的关系，修饰方向由空间词指向动作词）；

k. 方式（方式词和动作词之间的关系，方式词可为介、副词等词性，修饰方向由方式词指向动作词）；

l. 工具（工具词和动作词之间的关系，工具词可为介词、副词等词性，修饰方向由工具词指向动作词）；

m. 并列（目标词和并列词之间的关系，并列词通常为连词词性，修饰方向由并列词指向目标词）；

n. 陈述状态（状态词和述态词之间的关系，状态词可为名、介、形容词等词性，修饰方向由状态词指向述态词）；

o. 动变对象（对象词和动作、变化词之间的关系，修饰方向由对象词指向动作、变化词）；

p. 介词对象（对象词和作为时间、空间、方式、工具、陈述状态等关系中介词之间的关系，由对象词指向介词）。

2 基于修饰关系的语义分析

2.1 语句分析概述

语句是自然语言中最重要的结构单位，语句分析是自然语言处理的核心内容，语句分析最终是为了确定语句的语义，以达到理解语句的目的。语句语义是指句中每个词的意义及词之间的相互关系。语句分析通常的做法是按照词法分析、语法分析和语义分析的步骤来处理。词法分析的任务是把句中的词分离出来，并给每个词指派一个合适的词性；语法分析的任务是在词法分析的基础上进一步确定句法成分，如名词短语、动词短语、小句等，然后判断每个短语的句法功能，如主语、谓语、宾语等；语义分析的任务是确定语义角色，并最终得到句子的意义表示。然而这种做法中的每一个步骤都是困难的。在词法分析阶段，对于汉语这种词语之间无间隔标识的语言来说，要做到准确分词往往期待得到后面阶段的分析结果，对于词性确定，由于各种语言一词多性的现象很普遍，这种情况下必须考虑具体的语境才能确定某一词的词性，这意味着同样需要使用后面阶段的分析结果；在语法分析阶段，仅仅依靠词性的知识不足以确定短语的句法功能，文本的不规范性增加了分析的难度。正因为这样，较近的研究越来越重视对词汇特征的描写，通过在词汇语义和搭配中提取规则来克服语句分析中的困难。然而词汇规则和语句结构规则往往无限制地增加了系统的复杂程度。

2.2 基于修饰关系的语句分析框架

按照真实句子中词之间的修饰关系建立知识库后，对自然语言语句的分析过程可用图 1 表示。

在该分析框架下，修饰关系知识库和构句法是基础。通过词在修饰关系知识库中所处的位置可以确定语句中单个词的含义；通过对知识库中已有的修饰关系的计算可以确定语

句中词和词之间的修饰关系。由于假设待分析的语句是符合构句法的，因此在分析过程中可以按照构句法对语句逐步归约，以减少词之间修饰关系的判断次数。

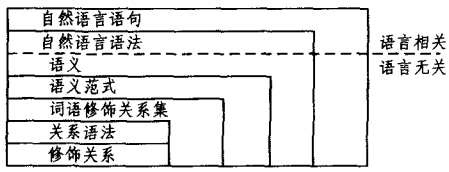

图 1　基于修饰关系的语言分析框架

2.2.1　构句法

构句法是指在语言形成过程中约定俗成的将句子成分连成合法语句的规则。语言学已总结出的构句法则通常是从词性角度来考虑的。下面给出的按照概念之间的修饰关系为基础的构句法是一种词序规则。由于在句子中的一个语法成分不能同时修饰多个另外的语法成分，因此根据修饰关系集和词序可以构成预期语义的句子。不同的语言词序略有差别，下面是基于汉语修饰关系的构句规则：

若 A 修饰 B，修饰关系为：

a. 动作、变化、述态、并列、动变对象、陈述状态、介词对象等，A 置于 B 后；

b. 情态、修饰、限定、度量、数目、时间、方式、工具等，A 置于 B 前；

c. 空间修饰关系，A 可置于 B 前或置于 B 后，按照置前处理；

d. 在同一语句中一个词可以被多个词修饰，一个词只能修饰一个另外的词；

e. 同一词既是修饰词同时也是被修饰词时，优先靠近其修饰词；

f. 被修饰词有若干前置修饰词时，按照修饰、限定、方式、工具、空间、时间、数目、度量、情态等由近至远的顺序排列。

2.2.2　语义范式

自然语言中存在一些词如“虽然”、“因为”等，连接一个句子，不充当句子成分，但有明确固定的语义。另外一些词如形容词、副词尾“的”、“地”或有语义的标点符号，在句中起调整语句结构和辅助语义作用。在分析包含这些词或符号的句子时，首先将该类词进行语义标记后从原句中略去，然后再进行语句分析。

2.3　基于修饰关系的语句分析过程

基于修饰关系的语句分析由以下几个步骤完成。

构建知识库。修饰关系知识库是对语句进行分析的基础，本文通过对真实文本语料库中的语句进行指导分析，将分析得到的修饰关系保存到知识库中。

分析预处理。该过程主要完成两部分任务：一是分词（将待句子中的词语用空格分开），二是按照预先定义好的语义范式对句子进行语义标记。从理论上，本文提出的分析方法在修饰关系足够多的情况下不需要预先进行分词处理。在做测试时，分词工作是使用北京大学语言研究所的分词系统结合部分手工操作完成的。

句子：国务院　总理　温家宝　主持　国务院　常委　会议　研究　太湖　流域　水　治理。

计算句子语义量。从左向右扫描句子，依次按照构句法确定句中的需要计算的两词，并根据知识库中的修饰关系计算两词之间是否有某种修饰关系，并确定其词语相关度。最后根据词语相关度计算句子语义量。图 2 显示了一个句子的基于修饰关系的语义分析树，表 1 是其对应的修饰关系集，同一个句子可能产生多棵语义分析树。

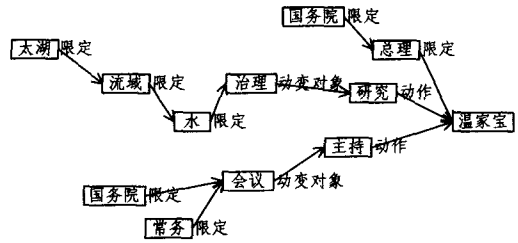

图 2　基于修饰关系的分析树

表 1　句子对应的修饰关系集

Index	Modifiers		HeadWords		Relation
	Index	DescWord	Index	ObjWord	
1	1	国务院	2	总理	限定
2	2	总理	3	温家宝	限定
3	4	主持	3	温家宝	动作
4	5	国务院	7	会议	限定
5	6	常务	7	会议	描写
6	7	会议	4	主持	动变对象
7	8	研究	3	温家宝	动作
8	9	太湖	10	流域	限定
9	10	流域	11	水	限定
10	11	水	12	治理	限定
11	12	治理	8	研究	动变对象

确定分析结果。从多个可能的分析中选择语义量最小的分析作为最终分析结果。通过分析结果可以确定句子的主、谓、宾等常规成分。

更新关系库。将句子分析出的修饰关系增加到修饰关系库中，这样在分析的过程中关系库可以得到扩充。

3　语句分析算法

第 1 节提出了基于修饰关系对语句进行处理的框架，在该框架下，语义分析是指根据知识库中已有的修饰关系，计算出输入语句中包含的各种修饰关系。这样能否对自然语言语句进行成功分析，取决于分析者知识库是否有足够的知识。下面首先给出相关概念的定义和计算公式，然后给出语义分析算法。

定义 2（信宿）　信宿是具有语句分析能力的知识库。信宿是一个四元组 $\wp=(C,R,Rn,Rf)$，其中 $C=(c_1,c_2,\cdots,c_n)$ 是一个有限词语集；$R=(\langle c_i,c_j\rangle)$ 是 $\langle c_i,c_j\rangle$ 的集合，$\langle c_i,c_j\rangle$ 为 C 上的两词语之间的修饰关系，其中 $\{c_i\}$ 表示修饰词的集合，$\{c_j\}$ 表示被修饰词的集合；Rn 是修饰关系类型的集合；Rf 是 R 到 Rn 之间的映射。

定义 3（词语相关度）　词语相关度用来衡量两个词语之间具有修饰关系的程度。设 c_i,c_j 是给定的两个不同词语（$c_i\neq c_j$），则 c_i,c_j 之间的词语相关度记为 $ST(\langle c_i,c_j\rangle,Rf_{c_i,c_j})$。

为了计算 $ST(\langle c_i,c_j\rangle,Rf_{c_i,c_j})$，定义集合：

$RT(c)=\{d|\langle c,d\rangle\in R\}$

$RC(c)=\{d|\langle c,d\rangle\in R \text{ or } \langle d,c\rangle\in R\}$

$RV(c_m,c_n)=\{Rf\mid Rf=(Rf_{c_k,c_m}=Rf_{c_k,c_n}) \text{ or } Rf=(Rf_{c_m,c_k}=Rf_{c_n,c_k})\}$

令

$$E(\langle c_i,c_j\rangle,Rf_{c_i,c_j})=\begin{cases}0, & c_j\notin RT(c_i)\\ C, & c_j\in RT(c_i)\end{cases}\quad (C\text{为常数})$$

$$F_k(\langle c_i,c_j\rangle,Rf_{c_i,c_k})=\frac{\overline{RV(c_j,c_k)}}{f}C\quad (f,C\text{为常数},c_k\in RC(c_i))$$

,则

$$ST(\langle c_i,c_j\rangle,Rf_{c_i,c_j})=\max_{k=1}^{n}(E,F_k)$$

$$SE(S)=\begin{cases}\infty, & \overline{\overline{MO(R(S))-M(R(S))}}\neq 1 \text{ or } \overline{\overline{W(S)-M(R(S))}}\neq 1\\ 1/\sum_{i=1,j=1}^{n} ST(c_i,c_j), & \overline{\overline{MO(R(S))-M(R(S))}}=1 \text{ and } \overline{\overline{W(S)-M(R(S))}}=1\end{cases}$$

3.1 语义分析算法

3.1.1 算法描述

算法 1(语义分析算法)

S:按照待分析句子顺序排列的词语集合
c:S 中词语
R:修饰关系的集合
M:修饰词的集合
MO:被修饰词的集合
(1)I:=1
(2)WHILE I<=LENGTH(S) DO
(3)ret:= calcST(c(I),c(I+1));
(4)IF ret==4 THEN
(5)newParseTree;
(6)ENDIF
(7)IF ret==1 or ret==4 THEN
(8)IF c(I) in M THEN
(9)newParseTree;
(10)ENDIF
(11)recRelation;
(12)ENDIF
(13)IF ret==2 or ret==4 THEN
(14)IF c(I+1) in M THEN
(15)newParseTree;
(16)ENDIF
(17)recRelation;
(18)ENDIF
(19)I:=I+1;
(20)ENDWHILE
(21)calcSE;

3.1.2 算法说明

calcST(a,b)函数根据知识库计算 a,b 词之间是否有修饰关系,计算结果可能有 4 种情况:(1)a 修饰 b;(2)b 修饰 a;(3)a,b 之间无修饰关系;(4)a 修饰 b 或者 b 修饰 a。对于(4),需要在原分析树基础上新增一棵分析树分别对应 a 修饰 b 和 b 修饰 a 的情况。

如果 c 为修饰词,且 c 已经在 M 中,由于同一词不能同时修饰多个词,因此需要新增一棵分析树。

NewParseTree 新增一棵分析树,该分析树复制部分原分析结果。

RecRelation 将分析结果记录到 R 中,不同的分析树对应不同的 R。

CalcSE 依次计算每棵分析树对应的句子语义量,选择语义量最小的分析树作为最终的分析结果。

3.2 语句构造算法

3.2.1 算法描述

定义 4(句子语义量) 句子语义量用来衡量语句的可理解程度。给定语句 S,该语句相对信宿 $\wp$ 的语义量定义为 $SE(S)$。

设 $W(S)$是语句 S 中词语的集合,$R(S)$是语句 S 中修饰关系的集合,$M(R(S))$是 $R(S)$中修饰词的集合,$MO(R(S))$是被修饰词的集合,则

算法 2(语句构造算法-生成树算法)

R={⟨c_i,c_j⟩|i,j<=n}:修饰关系的集合
(1)SET T=EMPTY
(2)SET S=R
(3)WHILE S NOT EMPTY
(4)FETCH an element⟨c_k,c_l⟩ from S
(5)IF ⟨c_k,c_l⟩ not IN T
(6)ADD ⟨c_k,c_l⟩ TO T according to sentence constructing rule
(7)DELETE ⟨c_k,c_l⟩ from S
(8)END WHILE

3.2.2 算法说明

算法根据句子语义关系表和构句法恢复原语句,是语义分析的逆过程,表明了语句、分析树和语义关系表之间的相互转换关系。

3.3 语句结构分析算法

3.3.1 算法描述

算法 3(语句结构分析算法)

S:语句词语集合
R:修饰关系的集合
M:R 中修饰词的集合
MO:R 中被修饰词的集合
SW:句子主语集;PW:句子谓语集;OW:句子宾语集
(1)SW= EMPTY;PW= EMPTY;OW= EMPTY
(2)WHILE S NOT EMPTY
(3)FETCH an element c TO SW from S
(4)IF c not in M THEN
(5)ADD c TO SW
(6)ENDWHILE
(7)FOR first element to last element of R
(8)FETCH ⟨c,d⟩ from R
(9)IF d in SW and (Rf='动作'or Rf='变化' or Rf='述态') THEN
(10)ADD c TO PW
(11)ENDFOR
(12) FOR first element to last element of R
(13)FETCH ⟨c,d⟩ from R
(14)IF d in PW and (Rf='动变对象' or Rf='陈述状态') THEN
(15)ADD c TO OW
(16)ENDFOR

3.3.2 算法说明

该算法根据语句语义关系表确定句子的主、谓、宾等主要成分,进而可以确定句子的其它成分结构。

4 测试及结果分析

为了评价基于修饰关系语义分析方法的有效性,使用了

中文自然语言文本语料来进行测试。当前作为自然语言语义处理研究用的语料库主要有语义依存网络语料库、人民日报语料库、汉语句法树库、国家语委现代汉语通用平衡语料库等。本文所需的训练、测试语料只需进行分词，不必预先标注。语料主要有两个来源：人民日报语料库和从人民网、新华网摘取的部分完整语句。为了便于训练，将题材限定在社会、政治方面。题材集中有利于使系统获得较深入的知识(修饰关系的稠密度较高)，这样可以在保证测试有效性的基础上控制训练的规模。

测试的目的是为了验证在已知修饰关系的基础上，系统能否对未知的修饰关系做出正确判断。在训练过程中选择了3个监测点，分别是录入至100个语句(1854个关系)、录入至300个语句(4636个关系)和录入至500个语句(7723个关系)时，记录对包含10个语句的测试集进行测试的结果。

下面首先给出两个性能评价指标。

定义5(关系判断正确率 RDAR)

$$RDAR=\frac{\text{判断正确的关系数}}{\text{总关系数}}$$

定义6(关系判断召回率 RDRR)

$$RDRR=\frac{\text{判断正确的关系数}}{\text{实际关系数}}$$

表2显示了在3个不同监测点下，测试得到的关系判断正确率和召回率。

表2 不同监测点下的关系判断正确率和召回率

Relations in KB	RDAR(%)	RDRR(%)
1854	40.0	18.2
4636	48.4	38.3
7723	60.1	69.5

从表2可以看出，关系判断正确率和召回率随着知识库中关系的增加，呈明显上升趋势。由于当关系数较少时，得到总的关系数比句子中实际的关系数要少，造成了第一行和第二行中正确率比召回率高的情况。

表3显示了3种情况下，根据得到关系的两种方式：直接在知识库中匹配(DC)和通过相关性计算(RC)的正确率和召回率。

表3 直接匹配和相关性计算的正确率和召回率

Relations in KB	RDAR		RDRR	
	DC(%)	RC(%)	DC(%)	RC(%)
1854	38.5	44.4	13.0	5.2
4636	45.7	57.1	27.9	10.4
7723	63.7	53.8	46.8	22.7

从表3可以看出，对于关系判断召回率来讲，直接匹配得到关系的方式起的作用大于通过相关性计算的方式；而对于关系判断正确率，前面两行相关性计算得到的关系准确率比直接匹配正确率高，第三行直接匹配的正确率比相关性计算方式高，这种情况表明了当知识库中关系数增加到一定程度时可能会对相关性计算产生噪声干扰。

对于采用修饰关系进行语义分析的系统来讲，较高的关系判断召回率是重要的。语言的构句法保证了语句结构的完整，句中某些词之间具有修饰关系但整体不能构成完整句子的分析会被废弃。

结束语 本文提出了一种基于词语之间修饰关系建立知识库，并用来指导自然语言句子语义分析的方法。设计了根据已有修饰关系判断未知关系的算法。确定了词语之间的修饰关系的类型并提出了依照修饰关系建立自然语言构句法的思路，在此基础建立了汉语依照修饰关系的构句法。提出了根据句子修饰关系表生成句子、根据句子修饰关系表确定句子成分的算法。根据该理论实现了汉语的语义分析系统并在该系统上完成了本文的实验部分。实验结果证明了对未知修饰关系判断的召回率和准确率随着知识库中修饰关系的增加而增加。

目前，相关方面的研究工作正在进一步开展。下一步的工作重点主要在以下几方面：(1)继续增加知识库的容量，研究基于修饰关系的类别算法；(2)完善和改进基于汉语的语义范式；(3)建立基于英语的修饰关系库和构句法；(4)探索基于修饰关系的英汉互译系统。

参考文献

[1] Fellbaum C. WordNet: An Electronic Lexical Database [M]. Cambridge: Mass MIT Press, 1999

[2] Hearst M A. Automated Discovery of WordNet Relations[M]// C. Fellbaum, ed. WordNet: An Electronic Lexical Database. Cambridge: Mass MIT Press, 1999

[3] Ruiz-Casado M, Alfonseca E. Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia[J]. Data & Knowledge Engineering, 2007, 61(3): 484-499

[4] Pedersen T, Pakhomov S. Measures of semantic similarity and relatedness in the biomedical domain[J]. Journal of Biomedical Informatics, 2007, 40(3): 288-299

[5] Terol R, Martínez-Barco P. A knowledge based method for the medical question answering problem[J]. Computers in Biology and Medicine, 2007, 37(10): 1511-1521

[6] Lee S, Huh S-Y. Automatic generation of concept hierarchies using WordNet[J]. Expert Systems with Applications, 2008, 35(3): 1132-1144

[7] Gomez F, Segami C. Semantic interpretation and knowledge extraction[J]. Knowledge-Based Systems, 2007, 20(1): 51-60

[8] Ruppenhofer J, Ellsworth M. FrameNet II : Extended Theory and Practice[OL]. http://framenet. icsi. berkeley. edu/book/book. html, 2006

[9] Hans B C. Bilingual FrameNet Dictionaries for Machine Translation[C]// M. GonzÃilez Rodrguez and C. Paz SuÃirez Araujo, eds. Proceedings of the Third International Conference on Language Resources and Evaluation. Las Palmas, Spain, 2002: 1364-1371

[10] Hans B C. Semantic Frames as Interlingual Representations for Multilingual Lexical Databases[J]. International Journal of Lexicography, 2005, 18(4): 445-478

[11] Nancy C, Narayanan S. Putting Frames in Perspective[C]// Proceedings of the Nineteenth International Conference on Computational Linguistics. Taipei, Taiwan, 2002

[12] Gacitua R, Sawyer P. A flexible framework to experiment with ontology learning techniques[J]. Knowledge-Based Systems, 2008, 21(3): 192-199

[13] Lee Chang-shing, Kao Yuan-fang. Automated ontology construction for unstructured text documents[J]. Data & Knowledge Engineering, 2007, 60(3): 547-566

(下转第290页)

个角度来说，TVO-MDP算法属于NC问题，那么NC=？P，这个问题上我们认为是相等的(也有人认为是不等的)，目前双方均没有找到相应的理论证明。在实际的时延测试向量排序过程中，我们进行多次实验，并对实验结果进行分析，得出若N值设定在$N \leqslant 2^8$这个合理的范围内，均可以取得良好的效果。

4 实验验证

实验在ISCAS'85，ISCAS'89平台上进行。选取了一些典型的电路。基于马尔可夫策略的新排序算法采用C++语言实现，所有的实验都在SUN工作站上运行，每个电路都使用零延时模型。测试序列由实验室的ATPG工具自动生成。完整的结果如表2所列，实验结果表明了TVO-MDP可以有效降低峰值功耗和平均功耗。

表2 时延测试向量排序的结果比较

电路名称	TVM	初始SA	最终SA	TVO-MDP优化效果
Rd73	128	21276	9517	55.27%
clip	167	45658	19241	57.86%
Sao2	125	15449	7379	52.24%
c5315	88	154257	136314	11.63%
c7552	118	266362	219246	17.69%
s420	64	5069	3605	28.88%
s510	70	10332	4678	54.72%
s820	142	39278	16558	57.84%
s832	144	40368	16948	58.02%
s1238	176	54464	29811	45.26%
s1488	151	79918	27004	66.21%

表2中TVM指时延测试向量数，SA指电路的开关活动数。从表2可以看到，采用TVO-MDP算法后，开关活动数平均降低了33.06%左右，最大的下降了66.21%。这些结果表明了TVO-MDP算法可以有效降低开关活动数，从而达到降低峰值功耗和平均功耗的目的。

为了进一步验证新方法TVO-MDP对测试功耗的有效性，也给出新方法与随机排序和优化排序的比较。其中的功耗用开关频率来表示，随机排序和优化排序的结果可以参考文献[3]的结果，最终结果如表3所列。

由表3可以看出，虽然对个别电路的优化效果并不明显，如S526，S444，但整体的优化效果还是令人满意的。从实验结果看，TVO-MDP方法特别适用于时延测试向量中不确定位较多的时延测试向量集。下一步的工作是要降低算法的时间复杂度，提高向量排序的优化准确度。

表3 TVO-MDP同随机和一般优化方法的比较(TVM：时延测试向量数)

电路	TVM	P_{random}	$P_{optimum}$	TVO-MDP
S838	748	120552	2318	2259
S832	980	178610	9864	9851
S820	964	170203	10137	10059
S713	427	54355	6395	6124
S641	404	49692	5460	5312
S526	556	48113	2239	2245
S510	443	52204	5462	5146
S444	365	25622	2381	2451
S420	364	30138	870	866
S400	356	21058	2273	2173
S344	258	21289	3221	3112

结束语 本文提出一种基于马尔可夫决策模型的时延测试向量排序新方法，它有效降低了测试功耗。算法在执行过程中需要进行转移概率的计算，因此在一定程度上增加了时间的消耗。这种测试时间的增加，相比于测试功耗的降低是值得的。从现在研究的发展来看，低测试功耗研究还需要做大量的工作。未来的工作主要集中在保证芯片质量和成本的前提下，进一步减少测试功耗。基于自动时延测试向量生成(ATPG)、可测性设计(DFT)和系统级的低功耗研究会继续得到加强。多种技术的优化整合也是亟待解决的问题。

参考文献

[1] 徐磊，孙义和，陈弘毅. 基于扫描的低测试功耗结构设计[J]. 计算机研究与发展，2001，38(12)：1423-1428

[2] 李晓维，李华伟，骆祖莹，等. 降低时延测试功耗的有效方法[J]. 计算机辅助设计与图形学学报，2002，14(8)：738-742

[3] Girard P，Guiller L，Landrault C，et al. A Test Vector Ordering Technique for Switching Activity Reduction during Test Operation[C]//Proceedings of Ninth Great Lakes Symposium. March 1999：24-27

[4] 韩银和，李晓维. 测试数据压缩和测试功耗协同优化技术[J]. 计算机辅助设计与图形学学报，2005，17(6)：1307-1311

[5] 向东，李开伟. 低成本的两级扫描测试结构[J]. 计算机学报，2006，29(5)：786-791

[6] 彭喜元，俞洋. 基于变游程编码的测度数据压缩算法[J]. 电子学报，2007，35(2)：197-201

[7] 王伟，韩银和，胡瑜，等. 一种有效的低扫描测试结构-PowerCut[J]. 计算机研究与发展，2007，44(3)：473-478

[8] 胡殿伟，向东. 采用时钟屏蔽策略降低测试功耗[J]. 清华大学学报：自然科学版，2007，47(7)：1216-1219

[9] 高阳，周如益，王皓，等. 平均奖赏强化学习算法研究[J]. 计算机学报，2007，30(8)：1372-1378

[10] Puterman M L. Markov decision proceses[M]. Hoboken：Wiley，2005：83-91

(上接第202页)

[14] Chomsky N. Syntactic Structures[M]. The Hague/Paris：Mouton，1957

[15] 杜小勇，李曼. 本体学习研究综述[J]. 软件学报，2006，17(9)：1837-1847

[16] 董振东，董强. 知网的理论发现[J]. 中文信息学报，2007，21(4)：3-9

[17] 孙景广，蔡东风. 基于知网的中文问题自动分类[J]. 中文信息学报，2007，21(1)：90-95

[18] 石晶，戴国忠. 基于知网的文本推理[J]. 中文信息学报，2006，20(1)：76-84

[19] 冯瑶，孙济庆. 一种基于知网的K-means聚类算法[J]. 情报学报，2007(3)

[20] 许云，樊孝忠. 基于知网的语义相关度计算[J]. 北京理工大学学报，2005，25(5)：411-414

[21] 林杏光. 短语结构语法[J]. 语言文字应用，1994(2)：58-64

[22] 赵军. 词汇功能语法[J]. 语言文字应用，1996(4)：104-108

[23] 苗传江，张庆旭. 功能合一语法[J]. 语言文字应用，1995(3)：76-81

[24] 张卫国. 广义短语结构语法述略[J]. 语言文字应用，1996(1)：73-79

[25] 吴云芳. HPSG理论简介[J]. 当代语言学，2003，5(3)：231-242，221