第16卷 第1期
2006年1月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol.16 No.1
Jan. 2006

基于改进ATN的句法分析

李海军,张 蕾
(西北大学 计算机系,陕西 西安 710069)

摘 要:句法分析是自然语言处理中的关键一环,目前的句法分析一般是依据句子中词的词性标记来进行的。而汉语单纯依据词性很难确定词之间正确的句法关系。在词类这个平面上进行句法分析存在着难以排除的结构歧义现象,因而使用语义知识排除结构歧义的方法更显重要。文中提出基于ATN(Augmented Translation Networks)的句法分析的新方法,对ATN进行相应的改进,利用《知网》的语义知识资源对其成分进行各种特性的标注。

关键词:句法分析;结构歧义;语义知识;标注;扩充转换网络
中图分类号:TP301.2 **文献标识码**:A **文章编号**:1005-3751(2006)01-0041-03

Syntactic Analysis Based on Improved ATN

LI Hai-jun, ZHANG Lei
(Department of Computer Science, Northwest University, Xi'an 710069, China)

Abstract: Syntactic analysis is the key in nature language processing. It is generally carried on according to the morphological feature marks of a sentence. But only using the morphological feature of a word, it's very difficult to confirm the correct syntactic relations between the words in a Chinese sentence. There are structure ambiguous phenomenon when analyes a Chinese sentence syntactically. It is very difficult to get rid of the structure ambiguities, if simply using morphological features of the words. Therefore it is important to use semantic knowledge to get rid of the structure ambiguities. This paper puts forward a new method based on improved ATN (Augmented Translation Networks). The method uses "Hownet" as the semantic knowledge resource to label the different composition in a Chinese sentence.

Key words: syntactic analysis; structure ambiguity; semantic knowledge; label; ATM

0 前 言

自然语言处理是人工智能一个重要的研究领域,而文本的句法分析在自然语言理解中占有重要的地位。在进行文本句法分析中,歧义的自动消解是一个关键而又尚未完全解决的问题。传统语法认为歧义包括词汇的多义歧义和结构的同形歧义[1]。歧义消解存在着基于"制约"和基于"优先"的方法。基于"制约"的方法是利用规则的形式来对分析进行约束排歧;而基于"优先"的方法则是从若干个已分析出的歧义候选结构中根据某种选择原则(preference)挑出一个最优的结构[2]。文中针对符合VP+N1+的+N2形式的句型进行分析,这种句型在单纯依靠语法来进行排歧时有时显得无能为力,例如维修图书馆的空调,可以得到两种不同的句法结构:((维修/VP+图书馆/N1)的+空调/N2)和(维修/VP+(图书馆/N1的+空调/N2)),这种结构对于人类的理解是没有什么问题的,但是对于计算机来说却会产生如上的歧义,如果仅从语法的角度来看,这两种都符合语法结构,要解决这类问题就需要有一定的语义知识辅助。

要正确理解一个句子不仅需要正确理解句子中各个词的词义,对其句法结构进行分析,而且必须考虑其语义结构。文中提出一种基于ATN的句法分析新方法,在对所分析出的每个成分进行特性标注时,对其加入《知网》的语义知识资源,用这种方法,可以同时解决词汇和结构上的歧义问题。

收稿日期:2005-04-01
基金项目:陕西省教育厅专项科研基金(HD01302)
作者简介:李海军(1977—),男,陕西蒲城人,硕士研究生,研究方向为人工智能与信息系统;张 蕾,博士,副教授,研究方向为人工智能与信息系统。

1 背景知识

1.1 扩充转换网络(ATN)

W.A.Woods在1970年提出的扩充转换网络是人工智能自然语言理解中的一种句法分析方法,ATN用递归语言类中的控制模型来识别各种语言类别,具有Turing机的威力[3,4]。ATN的概念来源于有限状态自动机,并在其状态转换图中增加了一些测试和动作。它可以把任意的输入文句转换成一棵带标记的树,从句法的角度清楚地描述了输入文句的结构。文中提出了在对其节点进行标注时加入《知网》的语义知识资源,由于在句子的结构和语义的共同层面上,语言具有一定的规律性[5],所以文中在对结构性歧义的排歧上面就是采用语义加语法的方法进行的。

1.2 知 网

《知网》是一个以汉语和英语的词语所代表的概念为描述对象,以揭示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[6]。《知网》中含有丰富的词汇语义知识和世界知识,为自然语言处理和机器翻译等方面的研究提供了宝贵的资源。语义词典是知网系统的基础文件。在这个文件中每一个词语义项的概念及其描述形成一个记录。目前词典中提供汉英双语的记录,每一种语言的每一个记录都主要包含 4 项内容。其中每一项都由两部分组成,中间以"="分隔。每一个"="的左侧是数据的域名,右侧是数据的值。它们排列如下:

NO. = 词或短语序号

W_X = 词或短语

G_X = 词或短语的词性

E_X = 词或短语的例子

DEF = 概念定义

E_X 中存放的是为那些具有多个义项提供的例子。这些例子强调例子的区别能力而不是它们的释义能力,用以为消除歧义提供可靠的帮助。除了语义词典外,知网还提供了义原分类树,分类树把各个义原及它们之间的联系以树的形式组织在一起,父子结点的义原具有上下位的关系。可以通过义原分类树计算义原间的语义距离。

知网的词典用文本的方式保存,要事先把它转换成方便系统实现的格式,鉴于系统是在 VC 的环境中实现,故采用 SQL Server 数据库对知网进行存储。

2 基本思想

利用 ATN 文法对句子进行分析,目的就是产生一个类似于寄存器样子的结构,这个结构描述整个句子,结构中的寄存器节点表示句子中的各种成分,形成一个以寄存器为节点的网络。每个节点是句子的不同语法成分,在对其分析过程中对每个成分标出它的各种特性。标注的越清楚,理解的也就越真切。在这里所说的特性就是指该句子的语气、语态、人称、时态、数、问题和类型等。图 1 就是一个语法成分的寄存器结构节点。

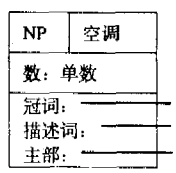

图 1 结构节点

节点中包含有一个语法类的成分和其特性,特性和成分都在节点中直接列出,并且各个成分用指针连到适当的其他节点。这些节点还可以再标上特性和成分,继续连接,直至终结节点。

词典是某种语言中所有词及其词类的表。其中的字(词)作为关键字,而相应的内容则为一组词类和一组特性向量,ATN 文法中所用的词典,其中的一些词还可以理解为有几个词义(wordsense),每个词义的词类可能是不同的,也可以有相同的词类和不同的特性。文中就是针对 ATN 分析过程中形成的寄存器节点所需要列出的成分和特性,在其中增加了该短语的语义信息和一定的语义角色,和图 1 中 NP 短语一样,在《知网》中对此短语的描述如下:

NO. = 050687

W_C = 空调

G_C = V

E_C =

DEF00 = 调整,PatientAttribute = 温度 DEF00 = 用具,* 调整,# 温度

DEF02 = {调整,PatientAttribute = 温度} DEF02 = {用具、* 调整,# 温度}

NO. = 050687

W_C = 空调

G_C = V

E_C =

把 NP 短语"空调"的语义成分加入其中的特性描述中,形成新的一个寄存器节点。如图 2 所示。

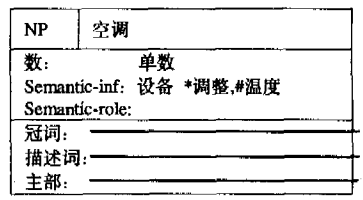

图 2 加入语义信息的结构节点

在节点中增加了 Semantic - inf 和 Semantic - role 项,其中 Semantic - inf 代表此 NP 短语的具体语义特征,Sementic - role 表示其在语义结构中可担当的语义角色,具体在《知网》中可以应用其 E_X 项来进行标注,文中就是在句法分析的基础上,加上其的语义特征用以排除句法分析的歧义。

3 基于改进 ATN 的句法分析模型

3.1 ATN 文法的构造

要有效地识别句子,必须加进一些知识或信息。ATN 是在递归转移网络的基础上扩充而成的,是加了标注的转换网络。ATN 所作的扩充:

·寄存器:存放信息,如单词、句子成分或其它的测试。

·测试:测试满足之后,才能通过某条弧。

·附加动作:通过某条弧时,这些动作被执行。

ATN 定义为按语法类组合在一起的结构模式,由一个标号、一组状态和节点组成的网络。网络中有一个状态为初始状态,并配有一组特性向量和成分名。网络中的弧描述某种转换,指定一个标号,开始于起始状态,终止于结束状态。ATN 在进行了一些附加信息的扩充之后,对句子的分析具有了很大的灵活性,但其依然存在着一些缺陷:

(1)当有歧义时网络中的回溯是不可避免的,系统的开销较大。

(2)对于结构性歧义的句子,可能产生几条不同的路径。

对于 ATN 所存在的缺陷,文中所采用的方法,能在某种程度上对于结构性歧义具有有效的排歧功能。

ATN 网络中的弧在网络中描述某种转换,所扩充的条件或动作有 4 种不同类型的弧:词类弧、查找弧、跳跃弧、发送弧,按照此构造 ATN 文法本身具有一定的语义特征[7]。再在具体的文本分析时利用《知网》的语义知识资源对其短语进行一定的语义特征标注。

3.2 《知网》语义知识的关联

利用 ATN 文法对文本进行分析,最终产生的是一种类似于寄存器结构的寄存器网络,结构中的寄存器节点表示句子中各种成分。为了对其充分的理解,分析过程中对每个语法成分节点都进行标注,文中就是利用《知网》强大的语义知识资源,来对其进行语义的标注,《知网》是面向计算机的,计算机化是《知网》的重要特色,《知网》作为一个知识系统,知识词典的常识性知识库是《知网》的最基本的数据库。《知网》的全部的主要文件包括知识词典构成了一个有机结合的知识系统。利用上述方法所构造的 ATN 文法对文本进行分析,对于有结构性歧义的句子,可能会有几种路径,这时就要对其进行语义层次上的理解,就要用《知网》的语义知识再次对其各个成分进行标注。

针对于 VP + N1 + 的 + N2 的这种结构,N1 可作为 VP 的宾语,述宾结构"VP + N1"加上"的"之后,作名词 N2 的定语,整个结构是一个定中结构,N1 又可与"的"结合在一起作 N2 的定语,"N1 + 的 + N2"这个名词词组再作为 VP 的宾语,整个结构就是一个述宾结构。按照 ATN 文法就可形成如图 3 的两棵句法树。

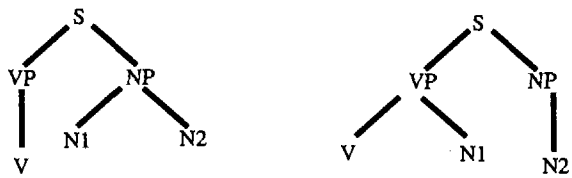

图 3　述宾和定中结构句法树

要让计算机分析出正确的句法结构,就要人工加入这种结构的一些规则,很显然这些规则是整个汉语短语结构规则集的一个子集,在利用规则的基础上还加上《知网》的语义信息,对其结构性的歧义进行排除。

对由 ATN 文法分析所形成的两种不同的句法结构进行评价要涉及到二个语义相似度的计算:义原间的语义相似度,词语间的语义相似度。而这二个相似度的计算就利用了《知网》的三个部分的知识:义原特征分类树、义项的概念定义。

词语相似度是一个主观性相当强的概念。脱离具体的应用去谈论词语相似度,很难得到一个统一的定义。因为词语之间的关系非常复杂,其相似或差异之处很难用一个简单的数值来进行度量。从某一角度看非常相似的词语,从另一个角度看,很可能差异非常大。

文中利用《知网》中的义原特征分类树来计算义原间的语义相似度。义原 a 与 b 的语义距离 Dis(a,b) = a 与 b 在义原分类树上的最短距离义原 a 与 b 的语义相似度。

$$Sim(a,b) = (1 - Dis(a,b)/(\text{义原分类树树高} * 2)) * 100$$

词语间的语义相似度利用刘群、李素建的相关思想计算[8],对计算所得值进行比较可以得出最优结果,如表 1 所示。

表 1　各词语语义相似度

	维修	图书馆	空调
维修	1	0.074074	0.285714
图书馆	0.074074	1	0.161932
空调	0.285714	0.161932	1

利用上面所阐述的方法进行各个相似度的计算,在取加权平均后得到各个短语的评价值,依据评价值的大小可以分析出正确的句法结构。根据以上值的比较可以得出:维修/vp 图书馆/n1 的空调/n2 的正确的句法结构应为(维修/vp(图书馆/n1 的空调/n2))。

4　实验结果分析

利用《知网》面向计算机的这个特点,把原来用文本存储的《知网》信息采用 SQL Server 数据库进行存储。由于文中要利用其中的义项实例集,再加上《知网》现在没有完全公开各个词语的实例集信息,所以对《知网》进行词语实例集的写入,使其语义信息得到进一步的完善。在利用以上的分析方法进行的句子分析中,采用 800 个 VP + N1 + 的 + N2 的句子作为试验语料。对这些句子先进行句法分析,得到的分析结果作为排歧系统的输入。排歧的正确率采用下面方法计算:

排歧的正确率 = 排歧后可得到正确分析结果的句子数/试验语料总句数

所得到的排歧结果的正确与否是通过手工进行判定的。目前试验结果排歧的正确率为 92.4%。针对于手工添加的语义实例,再加上在分析时对结果的正确与否采用手工进行判别,所以难免会有一些人为的因素在里面,为此,笔者会继续斟酌,力争其正确性的进一步提高。

5　结束语

文中提出的基于 ATN 的句法分析模型是在文本的句法分析中加入语义知识的一些尝试。它改进了 ATN 分析句子时对于结构性歧义所具有的缺陷,从而使这种文本处理的有力工具得到加强;建立了寄存器结构节点和《知网》知识描述语言的协同关系,为句法分析提供了语义分析的功能;利用了 ATN 表示方法和《知网》这种语义资源对汉语句法进行分析,解决了传统运用语法分析存在的结构歧义问题。试验结果表明这个模型对于汉语基本名词短语的分析是有效的。

参考文献:

[1]　邵敬敏.关于歧义结构的研讨:现代汉语语法研究的现状

(下转第 46 页)

列较好的实验结果[5,8,9]。

人们经过大量研究发现，蚂蚁个体之间是通过一种称之为外激素(pheromone)的物质进行信息传递，从而能相互协作，完成复杂的任务。蚂蚁在运动过程中，在它所经过的路径上留下该种物质，而且蚂蚁在运动过程中能够感知这种物质的存在及其强度，并以此指导自己的运动方向，蚂蚁倾向于朝着该物质强度高的方向移动，这样对于短的路径就会留下高浓度的外激素。因此，由大量蚂蚁组成的蚁群的集体行为便表现出一种信息正反馈现象：某一路径上走过的蚂蚁越多，则后来者选择该路径的概率就越大。蚂蚁个体之间就是通过这种信息的交流达到搜索食物的目的。在作业 $J_i(i=1,2,\cdots,m)$ 处分别设置 1 个蚂蚁，作业分配给处理机 j，蚂蚁就在处理机 j 上留下外激素，设 $\Gamma_j(j=1,2,\cdots,n)$ 表示处理机 j 的总的外激素，每个蚂蚁选择处理机 j 概率及外激素更新方程的为：

$$\begin{cases} P_j = \dfrac{\Gamma_j}{\sum_{j=1}^{n}\Gamma_j} \\ \Gamma_j^{new} = \rho\Gamma_j^{old} + \dfrac{Q}{F} \end{cases} \tag{2}$$

其中 P_j 代表蚂蚁选择处理机 j 的概率，F 为此次分配后完工时间，ρ 表示强度的持久性系数，一般取 0.15 ~ 0.9 左右，Q 为一正常数。

解分布式多处理机调度的蚁群算法如下：

(1) $l \leftarrow 0$(l 为循环次数)，给 $\Gamma_j(j=1,2,\cdots,n)$ 赋相同的数值，给出 ρ、Q 的值，随机给出一个调度方案；

(2) 对每个蚂蚁按转移概率 P_j 选择下一个节点，计算本次分配完工时间 F，按更新方程修改信息强度；

(3) 比较这次循环的结果，若目标函数 F 有改进，保留当前解为最好解，否则，外激素量采用上次最好解时的外激素量，$l \leftarrow l+1$；

(4) 若 $l >$ 规定的循环次数，记录当前蚂蚁的位置(当前的解)，停止运行，输出最好的解；否则转(2)。

在文中算法中，上述问题的受控赋时 Petri 网模型由程序按问题维数及其加工时间矩阵自动生成，其调度子网则在算法每次获得候选调度解后初始化。优化外环加工路径时令蚁群人工蚁数 $m=7$(即任务的个数)，选取外激素强度持久系数 $\rho=0.8$。可以得到表 4 所示的最优路径。

表 4 蚁群算法最优路径

	O_{00}	O_{01}	O_{02}	O_{10}	O_{11}	O_{12}
机器	D_0	D_3	D_2	D_2	D_1	D_3

并行传输中最大的花费时间(传输的代价)为 6，这与通过 Petri net 仿真得到的情况一致。

5 结束语

受控赋时 Petri net 模型和调度子网的规则，利用蚁群算法解决分布式资源任务调度，不会占用很多的系统时间，这样在解决大型模型时有较大优势，通过仿真试验可以选取较好的参数，可以更加节省机器时间。近年来，基于 Petri 网的复杂系统建模取得了较大的进展，Petri 网仿真评价将成为调度研究的一个有力工具。此外，蚁群优化作为一种新兴的离散优化方法解决 TSP、QAP 等组合优化问题都获得了非常好的结果，因而在分布式资源调度问题中引进蚂蚁算法是非常有前景的新方法。

参考文献：

[1] 袁崇义. Petri 网原理[M]. 北京：电子工业出版社，1997.

[2] 郑大钟，赵千川. 离散事件动态系统[M]. 北京：清华大学出版社，2000.

[3] 乐晓波，陈黎静. Petri 网应用综述[J]. 长沙交通学院学报，2004，20(2)：51－55.

[4] 蒋昌俊. 离散事件动态系统的 PN 机理论[M]. 北京：科技出版社，2000.

[5] 王笑蓉，吴铁军. 基于 Petri 网仿真的柔性生产调度[J]. 浙江大学学报(工学版)，2004，38(3)：286－291.

[6] Tanenbaum A S. 分布式操作系统[M]. 北京：电子工业出版社，1999.

[7] 孙 俊，须文波. 一种基于遗传算法的分布式系统的任务调度[J]. 计算机工程与应用，2003，39(21)：105－121.

[8] Dorigo M，Di Caro G. The Ant Colony Optomization: a new meta－heuristic[A]. in proceedings of the IEEE International Conference on Evolutionary Computation(ICEC 99)[C]. Piscataway，USA：IEEE Press，1470－1477.

[9] 许智宏，孙济洲. 基于蚂蚁算法的网格计算任务调度方法设计[J]. 天津大学学报，2004，37(5)：414－418.

(上接第 43 页)

和回顾[M]. 北京：语文出版社，1987. 218－229.

[2] 苑春法，黄锦辉. 基于语义知识的汉语句法结构排歧[J]. 中文信息学报，1999，13(1)：1－8.

[3] Winograd T. Language as a cognitive process[M]. Mas，USA：Addisom－Wesley Publishing Company. Inc，1983. 204－209.

[4] 顾跃挺. 机器翻译的语言模型与格语法[J]. 情报学报，1987，6(2)：118－120.

[5] 李卫东. ECAT 机译系统概述，语言与计算机(2)[M]. 北京：中国社会科学出版社，1985. 74－75.

[6] 董振东，董 强. 知网——知网简介[EB/OL]. http://www.keenage.com，1999.

[7] 姚天顺，朱靖波，扬 莹. 自然语言理解——种让机器懂得人类语言的研究(第 2 版)[M]. 北京：清华大学出版社，2002.

[8] 刘 群，李素建. 基于《知网》的词汇语义相似度计算[J]. Computational Linguistics and Chinese Language Processing，2002，7(2)：59－76.