ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
E-mail: eduf@cccc.net.cn
http://www.dnzs.net.cn

基于依存关系的中文句子语义分析研究

李华[1],朱敏[2]

(1.厦门大学 智能科学与技术系,福建 厦门 361005;2.福建省仿脑智能系统重点实验室(厦门大学),福建 厦门 361005)

摘要: 自然语言的机器理解是人工智能的一个重要的研究领域。为了挖掘自然语言中的语义关系,使计算机够像人一样去理解句子,该文使用哈工大语言技术平台的依存分析模块和知网及信息结构库,建立了一个语义分析系统来对于自然语言的依存结果进行处理。该系统实现了知网和LTP标记的一致化,并且建立了由信息结构库构建的信息模式树库,然后使用了嵌入式匹配以及基于树相似度和马科夫模型的词相似度算法来进行语义分析。通过实验,可以看到搭建的系统能够分析出句子的主要语义关系,系统对于语义分析的可行性得到了验证。

关键词: 依存; 中文; 语义; 知网; 自然语言

中图分类号:TP311　　文献标识码:A　　文章编号:1009-3044(2012)04-0856-04

Chinese Semantic Parsing Based on Dependency Relationship

LI Hua[1], ZHU Min[2]

(1.Cognitive Science Department, Xiamen University, Xiamen 361005,China;2.Fujian Key Laboratory of the Brain-like Intelligent System (Xiamen University), Xiamen 361005 ,China)

Abstract: Natural language understanding is an important field in Artificial Intelligence. To extract semantic relations from natural language, so that the computer can understand a sentence like human beings, a method of Chinese semantic parsing based on LTP and HowNet is presented. And the main architecture of the semantic parsing system is given out. Firstly, the markers of LTP are mapped to the markers of HowNet. Secondly, information structure model tree database is constructed by the information structures from HowNet. Thirdly, the result of dependency analysis is processed by using embedded tree matching method and word similarity computing using Markov Model and tree similarity. And the semantic relations are extracted by the semantic parsing system. Finally, the feasibility of the method is validated by experiments.

Key words: dependency;Chinese;semantic;HowNet;natural language

自然语言是在人类发展的过程中形成的,它的意义不仅仅是在于一种声音和符号,更是代表了人们说想要表达的更深层次的意义。这种存在于声音和符号背后的意义为人们之间的交流奠定了基础。自然语言的机器理解又称为计算语言学[4]。语义分析是自然语言理解的根本性问题,也是计算语言学研究中的重大难题[3],同时语义分析在机器翻译、问答系统、智能检索、语音识别等方面都有着重要的应用,所以研究语义分析具有重要的意义。计算机对于人类的自然语言进行语义学上的处理就称之为语义分析,计算机对于句子进行一系列的处理,根据分析出来的句法结构和其他各种信息,结合一定的知识库,能够对于句子所想要表达的意义用某种形式化的方法表示出来,使得能够将分析出来的结果用于后续的推理等过程。但是,中文句子的语义分析又和其他语言不一样,因为汉语具有其本身的特点:首先它是一种"意合"的语言[6],对于句子的形式没有严格的要求,所以使得对于汉语进行语义分析不能独立,也要结合"语形"来进行;其次汉语是一种比较灵活的语言,许多成分都可以省略,这使得计算机对于汉语进行处理没有一个通用的语境利用模型;再次在汉语的语义分析中要处理的歧义问题十分复杂,更增加了汉语语义分析的难度[4]。综上所述,汉语的语义分析十分重要,但是从目前来看,语义分析技术并不是十分成熟,相关的研究还待深入。

1 研究背景

计算机对于自然语言进行语义分析基于语义学理论。目前常用的语义学分析方法有概念依存理论,格语法,概念从属理论,语义场理论和知网等[2]。依存是一种将句子描写层级结构化的语言方法。依存语法最早是在法国语言学家泰尼埃的《结构句法基础》这一书当中提出的。他是公认的依存语法创始人。之后,1970年,美国语言学家罗宾孙(J. Robinson)在《依存结构和转换规则》中提出了关于依存语法的四条公理[5],为依存语法的形式化描述及在计算语言学中的应用奠定了基础。依存语法对于自然语言形式化的结果易于计算机处理,它在保留句子的短语结构信息的基础上直接表示出词和词之间的关系,对于进一步语义分析十分有利。依存语法认为动词作为中心词,其他的词受其支配,这样便于理清句子中词和词之间的关系。许多学者在中文语义依存结构方面都做了深入研究[10]。综上所述,我们认为采用基于依存关系的句法分析有利于进行句子的语义分析。

收稿日期:2011-12-06

作者简介:李华(1987-),男,福建福州人,厦门大学信息科学与技术学院硕士研究生,主要研究方向为自然语言处理;朱敏(1983-),女,内蒙古临河人,厦门大学哲学系博士研究生,主要研究方向认知逻辑。

2 知网及系统信息结构模式树库的构建

知网[9]是由机器翻译专家董振东创立的一个知识网络系统。这套系统致力于创建更加完善的基于知识的系统。知网在语义词典和世界知识方面有很丰富的资源,为自然语言处理提供了宝贵的研究资源。知网使用一种描述性语言KDML来对于词进行描述。在知网中有两个比较重要的概念,一个是词的概念,另外一个是词的义原。词的概念是指一个词的含义,在知网中用概念来描述一个词的含义,这样一个词可能存在多个概念,也就是多个含义,这也和现实中一词多义这个现象是一致的。而义原是存在于一个词的概念当中的,它构成了一个词的含义,也是不可分割的最小的基本单位[7]。知网通过考察大约6000多个汉字,对1500多个义原总结了上下位,同义,反义,对义,属性-宿主,部分-整体,材料-成品,事件角色等多种的义原之间的语义关系。我们通过知网提供的义原文件,充分利用知网的上下位关系,把词语的概念分析成为概念树,提出了一种融合了马科夫方法和树相似度的方法对于词语的相似度进行计算,具体的相似度计算方法将在另外的论文中详尽说明。这也是本文对句子进行语义分析的一个重要的组成部分。

知网信息结构库[9]是知网的体系的一个重要延伸,我们使用的知网信息结构库包含了271种信息结构模式。并且每种模式都有相应的例子。知网信息结构库是从大量的真实的语料当中提取出来的,为我们进行语义分析提供了十分珍贵的资源。信息结构库主要由信息结构描述以及例子、信息结构的句法结构索引两个文件共同构成的。

```
5.3.4 特别事件的表示
5.3.4.1 SYN_S=N <-- V --> N
        SEM_S=(位置) [处所] <— (事件,状态,存在) —> [存现体] (万物)
        Query1: N1有什么?
        Answer1: N1 + V + N2
        例子: 桌上-有-两本书, 后面-有-车, 树上-有-两只鸟, 前面-没有-路,
              水中-有-许多鱼, 锅里-有-鸡蛋, 门口-有-狗, 屋里-没有-家具,
              车里-没-人, 车外-有-人, 山顶-有-积雪, 台上-没有一一点灯光,
              路上-有-两个行人, 房前-有-一条小河, 家里-没-人, 家里-有-人,
```

图1 知网信息结构库中的一个信息结构

图1中所示的是信息结构库中的一条记录,SYN_S表示句法分布式,也就是该信息结构的词语排列结构,这里表示该信息结构由“名词+动词+名词”构成的。SEM_S则描述了该句法分布式的语义信息,在中括号中间的词语表示所连接的两个词之间的语义关系,比如处所表示第一个名词和中间的动词的语义关系是“处所”。在小括号里的词语表示与上面句法分布式相对应的具体词的描述,比如“位置”表示上面SYN_S中第一个N所对应的词应当为表示位置的词。Query和Answer表示该信息结构模式说传达出的真正的语义信息,为问答系统提供帮助。比如,在该例当中,通过该信息结构模式传达出处所和存现体的信息。最后信息结构库还为我们提供了大量的例子,以供我们参考和使用。

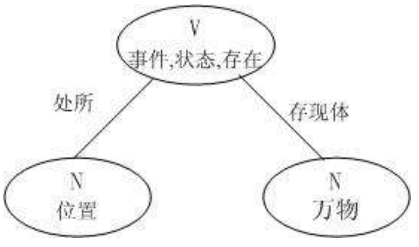

图2 信息结构模式树

在我们的实验当中,我们把上述的信息结构模式用我们定义的结构体InfoStructData存储,在这个结构体中除了包含上面的这些信息外,我们把SYN_S和SEM_S用树的方式进行存储。比如,上面的这个例子中我们把信息结构分析成为图2所示的树。在这棵树中每个节点存有该节点的词性,以及相对应的具体词,在树的边上存有两个节点之间的语义关系。我们通过这样的一棵树就可以清楚地表示出相对应的语义信息。这棵信息结构树也是我们下面工作的基础。我们按照上述方法将知网中的271个信息结构模式分析成为树状结构,构建信息结构模式树库用于后续的实验。

3 LTP平台及系统标注集映射

哈工大语言技术平台(Language Technology Platform,LTP)[8]是哈尔滨工业大学的社会计算与信息检索研究中心研发的一套系统。这个系统提供了一整套的自底向上的汉语语言处理模块,包括分词,词性标注,命名实体识别,依存句法分析,词义消歧以及语义角色标注等等。在2011年6月份该中心正式将LTP开源,在我们的实验中也使用了哈工大信息检索研究室语言技术平台中的依存句法分析这一模块。

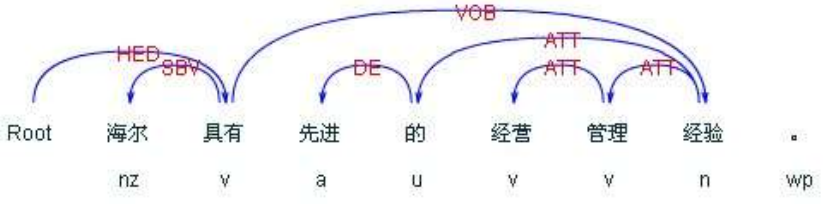

图3 LTP依存句法分析结果

在依存句法分析中,这一平台有着自己的标注体系。LTP的依存关系规范遵循语义原则和主干原则[11],选择语义上存在联系的

词语之间进行依存标注，并且选择句子中的主要的词作为依存关系的核心，其他的附属成分依存于核心词。比如句子“海尔具有先进的经营管理经验。”使用LTP分析结果如图3所示。这样我们通过哈工大这一平台就能够将句子分析成为依存树的形式，同时能够获得分词的结果和词性信息，这为下一步利用我们构建的信息结构模式树库进行语义分析提供了良好的基础。

但是由于哈工大语言技术平台采用的词性标注集与知网采用的标注集是两套不同的体系，所以我们在使用知网和信息结构模式树库对LTP平台分析的结果进行处理时需要进行标注集的对应。LTP使用的是863词性标注集。我们按照表1所示把两个标注集进行了映射（从863词性标注集映射到知网使用的标注集上）。知网中的PREFIX（前缀）和SUFFIX（后缀）INFSIGN（不定式符号）cha（汉字）PP（介词词组）AUX（助动词）这几个标记在信息结构库中并没有出现，故不进行映射。ws和x这两个LTP中的标记比较特殊，分别表示外文字符和非语素字，对于语义分析没有太大影响，所以进行特殊处理。

表1　知网与LTP标注集的对应关系

知网标记	LTP标记	知网标记	LTP标记
N	k, j, g, n, nd, nh, ni, nl, ns, nz, nt	ADV	d
NUM	m	PRON	r
CLAS	q	ECHO	o, e
V	v	ADJ	b, h
PREP	p	STRU	u
A	a	CONJ	c
PUNC	wp	EXPR	i
COOR	“又”		

需要说明的是STRU标记比较特殊，在信息结构库中有很多的词性为STRU的节点实际上对应LTP中的动词，所以在后续嵌入式匹配词性时遇到STRU词性的节点直接通过计算相似度来确定节点是否匹配。

4 系统搭建

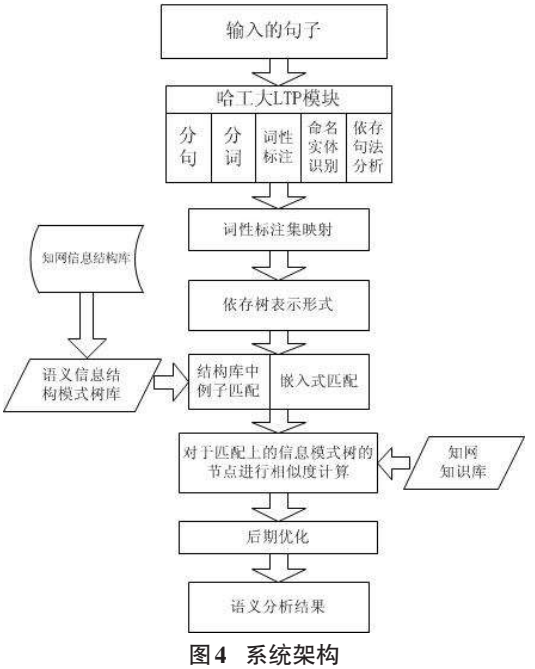

图4　系统架构

至此，我们已经完成了语义分析前的依存分析、语义信息结构模式树库的构建和相应的标记一致化工作。接下来，我们要进行基于依存关系的语义分析。图4是我们的系统的整体架构图，首先自然语言句子经过LTP分句、分词、词性标注、命名实体识别、依存句法分析之后进行我们上述的词性标注集的映射。然后转化成为依存树的表示形式。之后对信息结构库构造出来的信息结构模式树使用一种嵌入式匹配的方法[1]来对于依存树进行语义匹配。在语义信息结构模式树库中找到了词性标号相对应的信息结构模式树后还要对信息结构模式树的节点里的词语计算相似度，计算出的词语相似度范围为0到1之间，我们经过多次试验，设定了一个阈值0.75，当依存树中的节点和信息结构模式树中的节点的词语的相似度达到0.75时，我们认为两个节点是匹配的。如果嵌入匹配成功就在依存树的边上标注出对应的语义关系。在我们的系统中使用的嵌入式匹配算法是基于文献[1]提出的算法，下面我们将修改后的算法介绍如下：

首先以目标树的每一个节点为根构造子树，然后再调用嵌入节点匹配模块A对信息结构模式树库中的模式树ruleTree寻找与目标子树相匹配的节点，在匹配的过程当中我们不仅要对于节点的词性标号进行匹配，还调用我们的基于树相似度和马科夫模型

的词相似度匹配模块来对两个节点的词进行计算它们的相似度，当结果大于设定的阈值时，对应节点才真正匹配上，对于匹配上的节点记录在tempResult数组当中。FLAG标识了是否有嵌入匹配节点，如果有这样的嵌入匹配节点，则调用修正模块B:

```
For目标树的每一个节点targetNode为根节点的子树
        FLAG=true
        for ruleTree层次遍历每一个节点ruleNode
            模块A，
            若有嵌入匹配节点记录至tempResult
            否则：FLAG=false,退出
        end for
        if FLAG==true
            模块B
        Endif
end for
```

图5 嵌入式匹配主模块

```
For subTree的每一个节点snode
        在ruleTree中找到层次、词性相同节点node
        result=调用相似度模块计算词相似度（snode，node）
        If result>阈值&&node是根节点 then
            tempResult[node]=ruleNode标号
        Else
            If result>阈值&&tempResult[node->parent]包含ruleNode
                tempResult[node]里增加ruleNode标号
            Endif
        Endif
end for
If tempResult中有匹配的节点
        FLAG=true
Else
        FLAG=false
Endif
```

图6 模块A

1)在ruleTree当中寻找一个叶子节点leafNode。

2)在tempResult中找到一个包含leafNode的单元resultNode，在modifiedResult[resultNode]中添加leafNode，将resultNode置为目标树中resultNode号节点的父节点，leafNode也等于ruleTree中的父节点，继续上面的工作，直到找到ruleTree的根节点

3)ruleTree有无未处理叶子节点，若有，跳到1，若无，返回。

5 实验结果

由于语义分析没有固定的标准，所以我们采用人工判定的方法。在滨州中文树库中任选两句作为例子作为系统的输入，句子经过LTP的分词以及标注集的映射之后进行我们的语义分析，得到的输出结果如下：

第一句：为期/六/天/的/第五/届/北京/国际/图书/博览会/今天/在/北京/举行/。

第二句：中国/去年/发现/十/个/亿吨级/储量/规模/的/油气区。

表2 语义分析结果

第一句语义分析结果	第二句语义分析结果
内容(举行,博览会)	对象(发现,中国)
时间(今天 举行)	时间(发现,去年)
处所(在,北京)	对象(发现,油气区)
数量(博览会,届)	数量(油气区,个)
领属物(博览会,北京)	数量(个,十)
并列(博览会,图书)	“的“结构(规模,的)
整体(图书,国际)	限定(规模,亿吨级)
限定(北京,图书)	数量(级,吨)
数量(届,第五)	范围(规模,储量)
数量(天,六)	数量(吨,亿)
数量(为期,天)	标点符号(。)
标点符号(。)	

从这两个例子的分析结果我们可以看到，大部分的词和词之间的语义关系是正确的，并且较为全面和系统地对句子的主要意思进行了形式化的表述，说明这是一套有效的语义分析系统。这些分析出来的语义对推理和语言生成等后续处理具有重要的作用。同时由于信息结构模式中保存了问答信息，这对于将系统应用于问答系统也是非常方便的。 （下转第872页）

检测结果对比如表1所示。

表1 检测结果对比表

	实际个数	检测个数	错判个数	漏判个数	准确率	覆盖率
四分位法检测	59	256	217	20	15.23%	66.1%
四分位差距法检测	59	52	40	47	23.08%	20.34%
方差法检测	59	175	145	29	17.14%	50.85%
滑动四分位法检测	59	91	52	20	42.86%	66.1%
滑动四分位差距法检测	59	23	0	36	100%	38.98%
滑动方差检测	59	45	14	28	68.89%	52.54%

3 结束语

采用滑动窗口的分析方法对异常信号进行检测大大提高了对异常判别的准确率及覆盖率,其中滑动四分位差距检测方法的准确率达到100%。虽然覆盖率不高,但是可以通过加入小于1的检验系数进行调整,通过牺牲少许准确率扩大覆盖率。另外还可以根据时间序列的周期特征来定义滑动窗口的大小,从而达到更高的准确率及覆盖率。

参考文献:

[1] 张善文,雷英杰,冯有前.MATLAB在时间序列分析中的应[M],西安:西安电子科技大学出版社,2007:10-30 .
[2] 钟玉峰,雷国华.一种基于滑动窗口技术的入侵检测方法[J].信息技术,2009(7):166-170.
[3] 翁小清,沈钧毅.多变量时间序列异常样本的识别[J].模式识别与人工智能.2007.8(4):463-467.

(上接第859页)

6 结论和展望

语义分析是自然语言处理领域的难点和热点,知网和哈工大的语言技术平台为自然语言处理提供了宝贵的资源,本文在知网和信息结构库以及LTP的基础上,采用基于依存关系的语义分析方法,对句子依存分析的结果使用嵌入式匹配以及基于树相似度和马科夫模型相融合的词语相似度计算方法进行语义分析,实验结果表明这个系统能够对于句子进行有效的语义分析。但是知网和信息结构库还不是很完善,我们采用的方法还有需要改进的地方,比如没有考虑多个句子的语义组合等,下一步研究将在系统中引入FrameNet,以期能够更好地挖掘语义。

参考文献:

[1] 李剑锋.面向隐喻计算的汉语语义超常搭配识别模型研究[D].厦门:厦门大学,2008.
[2] 唐怡.用于常识推理的中文句子语义知识抽取[D].厦门:厦门大学,2010.
[3] 陈耀东,王挺,陈火旺.浅层语义分析研究[J].计算机研究与发展, 2008,45: 321-325.
[4] 湛志群,周昌乐.汉语机器理解研究现状及展望[J].电脑学习, 1999, 2: 3-5.
[5] 刘海涛.依存语法的理论与实践[M].北京:科学出版社,2009:7-11.
[6] 周昌乐.心脑计算举要[M].北京:清华大学出版社, 2003.
[7] Qiang Dong, Zhendong Dong. HowNet and the Computation of Meaning[M]. Singapore: World Scientific Publishing Company, 2006.
[8] 哈尔滨工业大学语言技术平台(Language Technology Platform,LTP)[EB/OL]. http://ir.hit.edu.cn/ltp/.
[9] 知网(HowNet)[EB/OL]. http://www.keenage.com/.
[10] YAN,J., D. B. Bracewell, F. REN and S. KUROIWA. A machine learning approach to determine semantic dependency structure in Chinese[Z]. Special Track at the Proceedings of the 19th International FLAIRS Conference, Melbourne Beach, FL, 2006, pp.782-786.
[11] 马金山.基于统计方法的汉语依存句法分析研究[D].哈尔滨:哈尔滨工业大学,2007.