基于依存语法的自然语言处理现状及前景展望

马永军
(襄樊学院 中文系,湖北 襄樊 441053)

[摘 要]依存语法强调动词中心说,注重句子成分之间的支配和被支配的关系,而不重视句子结构层次,易于自然语言处理中的句法分析系统和语义资源的构建。国内各科研机构自 90 年代以来逐渐开始大力发展基于依存语法的语料库的构建研究。把握依存语法的分析方法和原则,便于在自然语言处理中使用;开展依存语法的语料库建设,有利于依存语法在国内的研究得到发展。

[关键词]依存语法;语料库;标注
[中图分类号]H146 [文献标志码]A [文章编号]1000-8284(2007)10-0137-04

引言

在自然语言信息处理飞速发展的今天,面向真实语料的大规模语料库建设是当前最重要的任务之一。国家语委"十一五"的 6 个重点规划之一就是"语料库、知识库等语言工程建设"。而在语料库建设中,其重要的理论基础就是依存语法。当前国内外各大科研机构纷纷开发大规模语料库,他们的通常做法是:第一步,学习依存语法;第二步,确定标注的依存关系集,手工标注语料,建立标注语料库;第三步,机器自动学习,注意标注过程中的自动质量控制。鉴于依存语法在自然语言处理的语料库建设中的重要位置,笔者认为有必要对依存语法以及基于依存语法的自然语言处理做一个梳理,对当前的自然语言处理建设提点建议。

一、依存语法简介

1.1 缘起依存语法(Dependency Grammar)又称从属关系语法、配价语法,最早是 1959 年法国语言学家特斯尼耶尔(L. Tesnière)在《结构句法基础》(Element de Syntaxe Structurale)中提出的。其实他早在 1934 年就在他的论文《怎样建立一种句法》("Comment construire une syntaxe")中谈到了依存语法的基本论点。他被学者们公认为依存语法的创始人。

特斯尼耶尔未对依存语法下一个正面定义,目前比较通用的定义是周国光给配价语法下的定义:"一种结构语法。它主要研究以谓词为中心而构句时由深层语义结构映现为表层句法结构的状况及条件,谓词与体词之间的同现关系,并据此划分谓词的词类。"[1]

依存语法认为,一切结构句法现象可以概括为关联(connexion)、组合(jonction)和转位(translation)三大核心。句法关联建立起词与词之间的依存关系,这种依存关系是由支配词和被支配词联结而成的。它强调"动词中心说",认为动词是句子的中心,它支配着别的成分,而它本身则不受其它任何成分支配。直接受动词支配的有名词词组和副词词组,名词词组形成"行动元"(actant),副词词组形成"状态元"(circonstant)。

特斯尼耶尔使用图式(stemma)来表示动词、行动元和状态元,动词处于图式的顶部,以动词为分界线,行动元处于动词的左边,状态元处于动词的右边。图 1 是法语句子"Alfred parle bien"(Alfred 说得好)的图式:

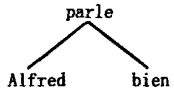


图 1 图式 stemma

特斯尼耶尔认为:从语义的观点看,第一个行动元就是行为的主体,即传统语法中的主语;第二个行动元是行为的目标,即传统语法中的直接补语(宾语);主语和宾语的区别是语义上的,而在结构上二者都是用来完善支配词的补语,主语和其他的补语没有什么不同;第三个行动元是行为的受益者或受害者,即传统语法中的间接补语(宾语);而状态元则相当于传统语法中的状语[2]。

在自然语言处理中,我们把表示依存关系的图示叫做依存树(Dependency tree,简称 D-tree),把表示短语结构关系的图示叫做短语树(Phrase tree,简称 P-tree)。这

[收稿日期]2007-05-22
[作者简介]马永军(1968-),男,湖北枣阳人,副教授,从事语文教育学、大学语文等课程教学研究。

两种方法是自然语言处理中最常用的分析方法。

图1就是一个依存树，在这个树形图中，结点上的标记都是单词，而不是非终结符，这样，我们就有可能使用依存语法直接表示句子中各个单词之间的依存关系，而不必使用标记，这种标记方法显然比短语结构语法更为直观，更为简洁，更便于计算机处理。

1.2 依存语法的分析方法和原则

依存语法其实只是一个抽象的概念，并不是一种严格定义的语法形式。它并未对句子分析明确地规定规则和方法，它强调句子中各个成分之间支配和被支配的关系、修饰和被修饰的关系，但是并未明确强调对每种依存关系加上标记。因此，很多学者都对依存语法的分析方法和转换原则进行了研究。

1960，美国语言学家海斯（Hays）提出“依存分析法”。1970，美国语言学家罗宾孙（Robinson）在《依存结构和转换规则》中提出四条公理，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，这四条公理是：1. 一个句子只有一个成分是独立的；2. 其他成分直接依存于某一成分；3. 任何一个成分都不能依存于两个或两个以上的成分；4. 如果A成分直接依存于B成分，而C成分在句子中位于A和B之间，那么C或者直接依存于A，或者直接依存于B，或者直接依存于A和B之间的某一成分。1987，美国语言学家舒伯尔特（Schubert）在此基础上提出了12条公理，他认为一个面向句法形式的、用于计算语言学应用的依存句法应符合以下基本原则：1. 句法只与语言符号的形式有关；2. 句法研究从词素到语篇各层次的形式特征；3. 词在句中通过依存关系相互关联；4. 依存关系是一种有向的同现关系；5. 词的句法形式通过词法、构词法和词序体现；6. 词对于其他词的句法功能通过依存关系来描述；7. 词组是一种作为一个整体与其他词和词组产生聚合关系的单位，它的成分间存在着句法关系，形成语言组合体；8. 一个语言组合体只有一个内部支配者，该支配者代表本语言组合体与句中其他成分发生联系；9. 除句子的主支配词外，句中的每一个词只有一个支配者；10. 每一个词只在依存结构中出现一次；11. 依存结构是一种真正的树结构；12. 应在依存结构中避免出现空结点。

罗宾孙的四条公理和舒伯尔特的12条公理为自然语言处理学界广为沿用。针对我国汉语的特点，我国的计算语言学家冯志伟先生根据机器翻译研究的实践，在他们的基础上提出了5条公理，认为一个D-tree应该满足如下5个条件：1. 单纯结点条件：在从属树中，只有终极结点，没有非终极结点，也就是说，从属树中的所有结点所代表的都是句子中实际出现的具体的单词。2. 单一父结点条件：在从属树中，除了根结点没有父结点之外，所有的结点都只有一个父结点。3. 独根结点条件：一个从属树只能有一个根结点，这个根结点，也就是从属树中惟一没有父结点的结点，这个根结点支配着其他的所有的结点。4. 非交条件：从属树中的树枝不能彼此相交。5. 互斥条件：从属树中的结点之间，从上到下的支配关系和从左到右的前于关系是互相排斥的，也就是说，如果两个结点之间存在着支配关系，那么，它们之间就不能存在前于关系[3]。

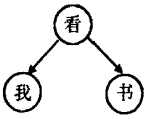
冯志伟先生提出的D-tree的这5个条件，更加形象地描述了从属树中各个结点之间的形式联系，进一步加深了我们对依存语法形式特性的认识，而且比罗宾孙的4条公理和舒伯尔特的12条原则更加直观，更加便于在自然语言处理中使用。

1.3 依存语法的形式化描述

在计算机上，句子语义依存关系的形式化描述为：

设：U = u1，u2，…，un 是一个长度为n的句子，U的语义依存关系表示为：SDL(U) = {SD(1)，SD(2)，…，SD(n)} 其中，SD(i) = (j，r)(j = 1，2，…，n)，第i个词是第j个词的修饰成分，语义关系为R，中心词为uj[4]。

一个典型的依存图如下：



作为自然语言处理的两大分析方法，依存语法比短语结构语法有着很多优势。例如，“铁路工人学习英语语法”这个句子，如果用短语结构语法来表示，其结构是一个短语结构树：

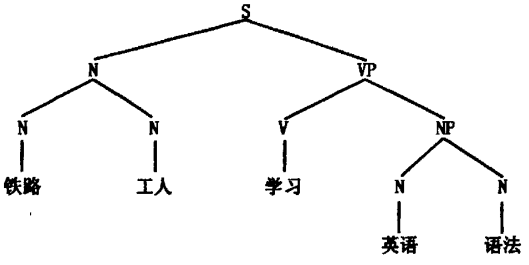


图2 短语结构树P-tree

经过转换后的依存树D-tree：

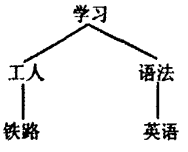


图3 转换得到的依存树D-tree

图2中，用短语结构语法分析后的句子有9个结点（node）、8条边，P-tree分析到最后是非终结符结点；图3中，用依存语法分析后的句子有5个结点、4条边，依存树上的所有结点都是句子中的词。很明显，图3比图2简洁很多，结点从9个减少到5个，在语言处理中，会方便很多[3]。

根据图2与图3，我们可以发现，与短语语法相比，依存语法具有以下四个特点[4]：

a）依存语法表示简单，不过多强调句子中固定的词

序；

b)依存语法中不含有非终极符结点；

c)易于表示句子的语义结构；

d)采用中心词驱动。

依存语法以动词中心，突出句子各成分之间的相互关系，支配和被支配、修饰和被修饰，而不必理会其他因素，分析方法简洁、明了。因此，目前国际上和国内，都在尝试运用基于依存语法的方法来处理自然语言。

二、依存语法在国内的发展

依存语法又称配价语法或从属关系语法，在语言理论界和计算语言学学界都有广泛运用，本文侧重探讨计算语言学学界的发展，顺带稍稍提及语言理论界的研究。

2.190 年代前

在语言理论界，上个世纪 70 年代末配价语法思想传入中国，朱德熙先生最早采用这种理论来研究汉语，他在《"的"字结构和判断句》(1978)中提到"单向动词"、"双向动词"、"三向动词"，指出可以把"向"的观念扩展到动词结构中去，并提出"歧义指数"理论[5]。张斌先生在《词语之间的搭配关系》(1982)中指出用向的概念来研究汉语语法现象的重要意义[6]。

而在计算语言学界，最早把依存语法引入自然语言计算机处理的应该是冯志伟先生。1980 年他在法国格勒诺布尔理科医科大学自动翻译中心 CETA 研究机器翻译时，基于依存语法设计出了汉—法/英/日/德/俄多语言自动翻译系统 FAJRA。

冯志伟先生在依存语法方面做出了巨大贡献，主要表现在两大方面：1. 他把特斯尼耶尔关于"价"的概念引入机器翻译的研究中，把动词和形容词的行动元分为主体者、对象者、受益者三个，把状态元分为时刻、时段、时间起点、时间终点、空间点、空间段、空间起点、空间终点、初态、末态、原因、结果、目的、工具、方式、范围、条件、作用、内容、论题、比较、伴随、程度、判断、陈述、附加、修饰等 27 个，以此来建立多语言的自动句法分析系统；对于一些表示观念、感情的名词，也分别给出了它们的价。2. 他还把依存语法和短语结构语法结合起来，在表示结构关系的树形图中，明确指出中心词的位置，并用核心(GOV)、枢轴(PIVOT)等结点来表示中心词。只要找出短语结构的中心词(GOV 或者 PIVOT)，就可以实现短语结构树和依存语法图式之间的转换[3]。

90 年代之前，国内学者们主要把依存语法应用于语言学理论方面，配价理论的研究稍稍起步，在自然语言处理学界，依存语法还没有绽放光彩。

2.290 年代后

90 年代以来，我国的中文信息处理研究者开始了解到依存语法，开始利用依存语法来进行汉语的自动处理，取得很好的成果，就笔者所知，清华大学计算机科学与技术系、北京大学计算语言学研究所、哈工大信息检索研究室、中国传媒大学的刘海涛、微软的黄昌宁、武汉大学语言与信息研究中心等等单位和科研机构都在进行基于依存语法的自然语言处理研究工作。从此，这种重要的语法才在我国语言信息处理界普及开来。

把依存语法与机器翻译结合起来，进行理论探讨的学者有很多，其中成就比较突出的一位就是中国传媒大学的刘海涛教授，他很早就开始介绍和研究依存语法，发表了许多相关论著，比较著名的就是 1997 年的《依存语法和机器翻译》一文，对依存语法在机器翻译中的应用做了很好的阐述[7]。

清华大学计算机科学与技术系在利用依存语法进行语料库的开发方面一直走在中国的前端。早在 90 年代初期，当我国基于语料库的汉语理解研究还处于初期探索阶段的时候，以黄昌宁教授为首的一批清华大学学者在这方面就作了有益的尝试。

他们首先采用依存语法对一个小型语料库进行人工依存语法标注，然后从这些语料库中抽取知识建立知识库，在这个知识库基础上对输入的句子进行自动句法分析，得到多棵可能的句法树，最后根据他们设计的一个评价函数得出正确的句法树。他们从中学地理课本中抽取的 500 条句子组成的语料库作了运行实验，在封闭语料中分析正确率达 90% 以上。另外，他们还结合语料库方法与规则方法的优点，设计了一个统计与规则并举的汉语句法分析模型 CRSP。在这个模型里，语料库用来支持各类知识和统计数据的获取，并检验句法分析的结果。规则主要用于邻接短语的合并和依存关系网的剪枝。他们的实验取得了令人满意的结果[8]。

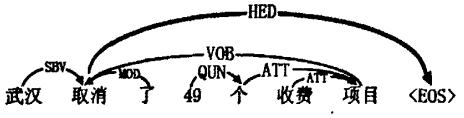
黄昌宁提出并论证了依存语法是合乎大规模真实文本处理要求的句法体系，并结合汉语的特点，研究了汉语的依存语法。他们最初把依存语法关系划分为 36 种依存关系，后来扩充到 106 种，最后又精简为 44 种[9]。

他们对待划分依存关系的态度变化恰恰反映了依存语法的特点。如果划分的太过粗泛，仅划分出"主谓宾、定状补"几种依存关系，就不能反映像"复数、限量词、连数"等句法关系；如果划分的太过细致，标注后的语料库就会规模过大，导致可操作性差，分析器的准确率也会受到影响。因此后来采取折衷的方法，现在他们的 44 种依存关系的划分受到许多学者的赞同和采纳。

哈工大信息检索研究室也是致力于基于依存语法的语料库建设的科研机构，做出了很多相关的项目和成果，如：2003 年的东芝中国研发中心项目"语料库加工"、2003 年的国家 863 子项目"基于依存分析的中文自动校对系统"、2007－2009 年的国家自然科学基金项目"汉语语义角色标注方法研究"。对外开放了一些免费资源，例如近年来建立的汉语依存树库[10]。

该汉语依存树库是一个开放资源，共有汉语句子 10000 句，每个句子占三行，第一行是经过分词和词性标注的句子，第二行在句子中每个词及词性的前面加上序

号,句子的末尾增加一个句尾标志“<EOS>”,由其支配全句的核心词,,第三行是句子中词与词之间的依存关系。依存关系中,每个关系以一个依存对表示,依存对中的第一个词是核心词,支配第二个词,如:“[2]公司_[1]我(ATT)”这个依存对表示“我”和“公司”存在依存关系ATT,其中,“公司”是这个关系的核心成分,“我”依存于“公司”。他们的标注规范共有34个依存关系类型。在该树库中,他们对汉语进行句法分析,将句子由一个线性序列转化为一棵结构化的依存分析树,通过依存弧反映句子中词汇之间的依存关系,例如:“武汉取消了49个收费项目”,依存分析的图示如下:



他们标注了6万句的大规模依存树库,提供较为丰富的词汇信息来源。然后通过对大规模依存树库的统计学习,获取其中的词汇依存信息,建立一个词汇化的概率分析模型。然后引入词汇支配度的概念。

另外还有许多大学的计算语言学处理中心都在致力于用依存语法来建立语料库标注系统,不过方法都大同小异,这里不再一一赘述。

三、展望

当前国家越来越重视基于依存语法或其他语法的语料库建设问题。2006年11月,在国家语委“十一五”科研工作会议总结上,国家语委副主任、教育部语言文字信息管理司司长李宇明指出,“十一五”期间,语言文字应用研究有六大重点课题,其中第四大课题就是“语料库、知识库等语言工程建设”,具体内容为:研制语料库、知识库等的建库规范及其语言文字加工规范,研究相关的知识产权问题,推进语料库、知识库的社会应用与共享。通过语言工程的建设来促进语言文字的规范化和信息化工作,促进包括少数民族语言在内的多语种信息平台建设,促进国家语言资源监测与研究中心的建设,促进用动态流通语料库等手段监测、研究语言生活状况。

在过去的2006年,语言学方面的国家社科基金项目,76个全国社科项目中有8个都是语料库项目。在今年2007年国家社科基金(语言学)项目的申报指南中,明确指出“当前国外计算语言学的显著特点是:提倡建立语料库,使用机器学习的方法获取语言知识;越来越多地使用统计数学方法分析语言数据;构造通用和专用的语料库。我国计算语言学研究应瞄准学科研究的方向,开展一些扎实的基础研究。比如,面向中文信息处理的语言文字研究,包括实用的词典、语义系统、语法规则、完备的语言知识体系等,特别要加强面向大规模真实文本的内容计算的语言知识的挖掘和形式表示等方面的研究。”

可见,利用依存语法,结合其他的格语法、题元理论,建立基于依存语法的词汇语义资源的方法在国家政策的大力支持下、在国内众多科研机构和学者的努力下、在国际提倡语料库建设的大形势下,一定会在不久的将来大放异彩。

[参 考 文 献]

[1] 周国光.汉语配价语法论略[J].南京师大学报(社科版),1994,(4).

[2] 冯志伟.现代语言学流派[M].陕西人民出版社,1999:160-170.

[3] 冯志伟.机器翻译研究[M].中国对外翻译出版公司,2004:431-434.

[4] 李涓子.基于语义依存关系的汉语理解模型研究[D].清华大学博士后出站报告,2001.

[5] 朱德熙.“的”字结构和判断句[J].中国语文,1978,(1).

[6] 文炼.词语之间的搭配关系[J].中国语文,1982,(1).

[7] 刘海涛.依存语法和机器翻译[J].语言文字应用,1997,(3).

[8] 谌志群,周昌乐.汉语机器理解研究现状及展望[J].电脑学习,1999,(2).

[9] 周明,黄昌宁.面向语料库标注的汉语依存体系的探讨[J].中文信息学报,1994,(8).

[10] 参见哈尔滨工业大学信息检索研究室主页:http://ir.hit.edu.cn.

[责任编辑:张正明]

基于依存语法的自然语言处理现状及前景展望

作者： 马永军
作者单位： 襄樊学院,中文系,湖北,襄樊,441053
刊名： 学术交流 PKU CSSCI
英文刊名： ACADEMIC EXCHANGE
年，卷(期)： 2007，(10)
被引用次数： 0次

参考文献(10条)

1. 朱德熙 "的"字结构和判断句 1978(01)

2. 李涓子 基于语义依存关系的汉语理解模型研究 2001

3. 冯志伟 机器翻译研究 2004

4. 冯志伟 现代语言学流派 1999

5. 周围光 汉语配价语法论略 1994(04)

6. 查看详情

7. 周明;黄昌宁 面向语料库标注的汉语依存体系的探讨 1994(08)

8. 谌志群;周昌乐 汉语机器理解研究现状及展望[期刊论文]-电脑学习 1999(02)

9. 刘海涛 依存语法和机器翻译 1997(03)

10. 文炼 词语之间的搭配关系 1982(01)

相似文献(10条)

1. 会议论文 陈波 基于依存语法的语料库标注研究 2007
依存关系的标注一直是近年来计算语言学界语言资源建设的的主流之一。本文从理论研究和实践研究两大方面对国内外依存语法标注的进展做了一个大致梳理,针对汉语依存语法标注的研究现状,提出了一些建议。

2. 学位论文 许伟 句法-语义一体化的汉语句法分析研究 1997
该文把词的语义分类引入了汉语句法分析.在具体词的搭配实例的基础上通过义类标注建立了语义关联网.同时以人工进行依存语法句法标注过的语料库(树库)为资源,通过机器学习自动得到了三个规则集,即:上下文无关的二元规则集、禁止规则集和冲突规则集.以三个规则集为基础再结合语义关联网,作者建立了一个句法分析实验系统,并对近1000个句子(句长平均23个词)做了句法分析.实验结果证明了该文研究方法的可行性.

3. 会议论文 陈波 基于依存语法的自然语言处理现状及前景展望 2007
基于依存语法的大规模语料库建设是近年来的国家社科重大课题。本文对依存语法以及基于依存语法的语料库建设进行了梳理,依存语法强调动词中心说,注重句子成分之间的支配和被支配的关系,而不重视句子结构层次,易于自然语言处理中的句法分析系统和语义资源的构建.国内各科研机构自90年代以来逐渐开始大力发展基于依存语法的语料库的构建研究.

4. 学位论文 刘伟权 自然语言理解与汉语文本信息处理理论研究 1997
该文对自然语言理解的基础理论,特别是汉语文本信息处理的理论和应用作了系统的研究.由于自然语言理解是一门交叉性学科,研究工作宜从语言学、认知科学、信息论和情报学这四个方面同时展开.

5. 学位论文 熊德意 基于括号转录语法和依存语法的统计机器翻译研究 2007
基于句法的统计机器翻译近年来逐渐成为统计机器翻译的研究热点.基于句法的模型有助于解决基于短语的模型所面临的主要问题,如短语层次上的重排序,泛化能力弱,以及要求短语连续等问题.语言学意义上基于句法的模型,还可以将源语言端、目标语言端的语言学知识引入到翻译模型中,从而极大地改善译文的质量.
本文在基于句法的统计机器翻译框架下,针对短语模型的主要问题,在括号转录语法的基础上提出了基于最大熵的括号转录语法模型,在依存语法的基础上提出了依存treelet-string对应模型.为了支持基于依存语法模型的研究,本文在统计汉语句法分析方面也进行了深入的研究.在以上三个方面,取得了以下主要成果:1.多知识源驱动的统计汉语句法分析句法分析的准确度和速度对于基于句法的统计机器翻译来说至关重要.在提高句法分析准确度方面,本文采用多种策略,将树库内部和外部的知识融合到统计句法分析模型中.首先改造了中心词映射表,并对一些短语进行重标注,从而充分利用了树库内部的词汇知识和语法知识.其次构建了一个单词.类的选择偏向模型,将树库外部的语义知识引入到句法分析中,使句法分析F1值提高了0.9%,错误率下降了4.4%.进一步的错误分析表明语义知识在复合名词短语,并列结构以及名/动词性标注消歧方面都有很大作用.在提高句法分析速度方面,本文定义了两种估计量来近似估计边的外向概率:先验估计量和边界估计量.由这两种估计量构成的组合估计量使句法分析器在性能不变的情况下,速度提高了1.5倍.2.基于最大熵括号转录语法模型的统计机器翻译针对括号转录语法(BTG)模型没有提供一个机制来确定相邻语块顺序的缺陷,本文提出了基于最大熵的括号转录语法(Maximum Entropy Based BTG,下文简称为MEBTG)模型.该模型将BTG中预测相邻语块顺序问题看作是一个分类问题,从而引入最大熵分类器,构建最大熵重排序模型.本文提出了重排序实例抽取算法,同时将双语语块的边界单词作为最大熵的分类特征.总体上来说,最大熵重排序模型相对于其它重排序模型,如距离惩罚模型,先验概率模型,词汇化模型,具有诸多优点.它是基于特征的,因而具有一定的泛化能力;它是和内容相关的,并且采用判别式训练,因而充分利用了训练语料库中的信息;同时它也是层次化的,在一定程度上能够处理远距离重排序.本文在.MEBTG模型基础上实现了一个实际的翻译系统Bruin,系统的核心模块解码器是基于CYK算法设计的.实验表明,最大熵重排序模型显著地提高了系统的BLEU值.在大规模语料上,Bruin系统引进了一些新技术,性能获得了极大提高.这些技术包括建立双语言模型,以及引入重排序窗口和标点符号来限制重排序等.3.基于依存treelet-string对应模型的统计机器翻译为了将语言学知识集成到翻译模型中,本文提出了一个新的基于依存语法的模型:依存treelet-string对应(DTSC)模型.该模型将源语言的依存结构树映射到目标语言的串上.DTSC模型具有很强的灵活性和表达能力.它能够描述多层树结构,具有泛化能力,可以处理与中心词相关的不同结构的重排序问题,通过引入变量和间隔允许源语言目标语言两端的短语非连续,最后它可以与短语模型充分兼容.本文给出了DTSC的抽取算法,以及DTSC模型与N-

gram语言模型的融合方法.为DTSC模型设计了Chart风格的解码器算法,在算法中引入了两种基本操作:替换和粘接.在DTSC模型的基础上本文实现了一个翻译系统Mo-tse,给出了Mo-tse与Bruin的对比实验,以及译文结果分析.

6.期刊论文 苏菲.马翠霞.戴国忠 针对特定几何语言的句法语义一体化分析方法 -计算机工程与设计2004,25(10)

提出了一种句法语义一体化的语言分析方法,句法分析和语义理解时采用并行方法,利用两者之间的相互关系实现句法和语义的分析.针对自然语言理解在几何特定领域的约束性,以依存语法为基础,利用标注过的语料库知识,采用规则统计模型,对已经标注好词性语义的句子词串进行句法语义一体化分析,生成符合数学规范的数学表达式.实验证明,建立的系统对100个几何描述的句子进行测试,得到的正确率为98%,在几何领域具有良好的实用性,能够满足实际的需要.

7.学位论文 高玲玲 基于依存语法的汉语句法分析研究 2009

句法分析是自然语言处理基础研究中的一个关键技术之一，是衔接词法分析与语义分析的桥梁。本文的目的是从汉语自身特点出发，以现有的句法分析理论和方法为指导，研究和开发适合汉语的句法分析技术。

句法分析技术指的是依据语法规则来确定句子结构的分析方法。依存语法是当今句法学研究的前沿和热点问题之一，本文的句法分析采用的语法体系就是依存语法，采用的句法分析技术是决策式依存句法分析方法。Niver算法作为决策式句法分析方法已经成功的应用于英文的依存句法分析，因为英文和中文在句法特点上具有一定的相似性，所以本文采用Nivre算法进行汉语依存句法分析。

本文首先对现有的一些依存句法分析方法从处理策略，算法的时间复杂度等方面进行了综合分析和比较，其中详细研究了Nivre算法，然后针对该算法，本文提出了进一步的改进。Nivre算法在分析长距离右依存时会出现错误，在汉语中，只有动词和介词跟他们的依存者具有右依存关系，所以错误主要发生在动词和介词的依存分析上。本文依据汉语介词短语的特点，提出了一种改进的Nivre算法，让除了介词外的介词短语部分先进行依存分析，最后再是跟介词之间的依存分析，来减少介词的长距离依存问题，提高汉语依存句法分析的正确率。

实验数据采用含有1万个句子的哈尔滨工业大学的依存关系语料库，采用基于支持向量机(SVM)的句法分析器MaltParser作为本文算法的实现工具。结果表明，使用改进后的Niver算法进行汉语依存句法分析，正确率提高了1.72%，对介词的长距离依存取得了比较好的分析结果。

8.期刊论文 李明琴.李涓子.王作英.陆大(乡 金).LI Ming-Qin.LI Juan-zi.WANG Zuo-ying.LU Da-Jin 中文语义依存关系分析的统计模型 -计算机学报2004,27(12)

该文提出了一个统计语义分析器,它能够发现中文句子中的语义依存关系.这些语义依存关系可以用于表示句子的意义和结构.语义分析器在1百万词的标有语义依存关系的语料库(语义依存网络语料库,SDN)上训练并测试,文中设计、实现了多个实验以分析语义分析器的性能.实验结果表明,分析器在非限定领域中表现出了较好的性能,分析正确率与中文句法分析器基本相当.

9.学位论文 杨军玲 汉语动词词语搭配自动获取方法研究 2006

在自然语言处理领域中，句法分析是实现语言“理解”的必然环节，也是公认的一个重点和难点。面向依存文法的句法分析方法主要是通过获取句子的核心动词及其所支配的词语搭配，进而分析句子内词语之间的依存关系，以建立依存句法树。依存语法认为动词是句子的中心，动词在汉语句子中起支配作用，因此动词组合框架的研究能为自动句法分析和处理提供较好的基础。

本文主要基于语料库的动词词语搭配自动获取方法进行了研究和实验。

(1)由于目标动词和搭配词的词性标注在搭配获取工作中占有重要地位，因此，作为基础性工作，我们首先研究了词性标注中兼类词的排歧问题。利用粗糙集约简理论提出了一种基于非完备决策表的兼类词标注校对规则获取方法，以作为基于软件的词性标注结果的辅助校对工具，旨在提高兼类词词性标注的正确率，为获得高质量的语料库提供基础。

(2)在确保高质量语料库基础的前提下，探讨了面向依存语法分析的动宾搭配自动获取的方法。通过对已有方法的概括，在词语搭配上重点研究并分析了互信息、Cosine系数、x2测试、似然比4种较优的词语度量方法，比较了方差、离散度、熵3种结构度量方法优劣。随后提出了一种基于互信息和信息熵融合的搭配获取方法，将其应用到动名、动动搭配的获取，在高频下取得了较好的效果。

(3)首次将最大熵模型应用于动词词语搭配的获取。以动动搭配的获取为着眼点，抽取搭配词对的上下文词性信息及其关联程度的统计信息构造候选复合特征模板，结合粗糙集理论的约简技术，获得训练最大熵模型的最简特征模板。一系列实验证明，基于最大熵模型的动动搭配的获取方法是可行的。

最后，对动词词语搭配获取的未来研究进行了展望。

10.期刊论文 赵军.黄昌宁.ZHAO Jun.HUANG Chang-ning 汉语基本名词短语结构分析模型 -计算机学报1999,22(2)

本文提出了用词语潜在依存关系分析汉语baseNP结构的模型,它有以下的特点:①将依存语法知识融入概率模型中,使得baseNP结构分析在依存语法知识的指导下进行,其性能优于纯粹的概率模型--相邻模型;②词语潜在依存强度的获取算法是基于MDL原则的,在模型建造时既考虑数据拟合性,又考虑模型归纳性,其性能优于基于极大似然原则的词语潜在依存强度获取算法;③词语潜在依存强度获取算法在复杂特征集上进行,可以有效地解决参数估计中的数据稀疏问题.实验结果显示,这个模型对于汉语baseNP结构分析是有效的.