第50卷第4期
2010年7月

大连理工大学学报
Journal of Dalian University of Technology

Vol. 50, No. 4
July 2010

文章编号：1000-8608(2010)04-0609-05

短语结构到依存结构树库转换研究

周惠巍*， 黄德根， 钱志强， 杨元生

（大连理工大学 计算机科学与技术学院，辽宁 大连 116024）

摘要： 为基于真实语料进行句法分析，构建了大规模的短语结构树库和依存结构树库，并尝试在两种结构的树库之间进行转换. 讨论了宾州中文树库(Penn Chinese Treebank，CTB)中短语结构树库和依存结构树库的关系，并基于现代中文依存文法制定了中心子节点过滤表，依据该表将短语结构的CTB转换为依存结构树库. 在CTB中随机抽取200句语料，转换正确率达到了99.50%. 基于该转换得到的依存结构树库可以进一步进行中文依存关系解析的研究.

关键词： 句法分析；短语结构；依存结构；树库；转换

中图分类号： TP18 **文献标志码：** A

0 引 言

汉语句法分析一直被认为是中文信息处理的一个重要技术. 其可以分为两种方法：一种是短语结构解析，即将句子划分成短语，解析句子短语间的层次结构关系[1]；另一种是依存结构解析，即解析句子词语间的依存关系[2,3]. 依存关系可以明确地表明词语间的支配关系，并能方便地发展为语义依存描述，近年来越来越受到研究人员的重视.

计算语言学的发展使得人们可以使用机器学习的方法基于大规模的语料库构建汉语句法解析器. 国内外研究人员已经构建完成了各种语言的树库，树库标注方式各不相同，其结构大体分为两类：短语结构树库和依存结构树库. 常用的汉语树库有短语结构的宾州中文树库(Penn Chinese Treebank, CTB)[4] 和依存结构的哈尔滨工业大学信息检索研究室的依存结构树库(HIT Chinese Dependency Treebank, HIT-IR-CDT). CTB的创建为汉语句法分析提供了一个平台，基于CTB的短语结构解析[5] 和依存结构解析[6] 成为当前研究的热点. 但是CTB是短语结构树库，而且没有标明每个短语的中心子节点，所以必须首先将CTB转换为依存结构树库，然后基于转换后的CTB才能进行中文依存关系解析.

目前，汉语树库资源仍然存在不足，而手工标注树库需要大量的时间和人力. 短语结构树库和依存结构树库虽然标注方法不同，但描述的都是句法结构，在深层次上具有一致性，所以人们尝试在两种不同类型的树库之间进行相互转换[7~9]. Lin提出基于中心子节点表(tree head table)将短语结构树转换为依存结构树的算法[7]. 转换采用递归算法，将短语结构树中各短语的非中心子节点依存到该短语的中心子节点上. Xia提出基于中心子节点过滤表(head percolation table)将宾州英文树库(English Penn Treebank，PTB)和CTB从短语结构树转换为依存结构树的算法[8]. 中心子节点过滤表与中心子节点表的概念基本相同，均由短语类型、搜索短语中心子节点的方向以及短语中各类标注的子节点作为中心子节点的优先次序三部分组成. 但是Xia构造的汉语中心子节点过滤表非常简单，表中每种短语包含的子节点类型比实际CTB语料中出现的子节点类型少许多. 比如名词短语(noun phrase，NP)的子节点类型为43种，而Xia构造的中心子节点过滤表的NP子节点类型仅为5种. 如果在短语中没有搜索到中心子节点过滤表中的节点类型，Xia即将搜索起始节点默认为该短语的中心子节点. 这种处理方法简单可行，但是需要完整的中心子节点过滤表，否则会影响转换效果. 党政法等提出了一种

收稿日期： 2008-08-09； **修回日期：** 2010-04-04.
基金项目： “八六三”国家高科技发展计划资助项目(2006AA012140).
作者简介： 周惠巍*(1969-)，女，讲师，E-mail：zhou_huiwei@163.com；黄德根(1965-)，男，博士，教授.

将清华汉语树库(Tsinghua Chinese Treebank,简称 TCT)从短语结构转换成依存结构的算法[9]. 该算法充分利用了 TCT 具有的功能、结构双重标记,转换得到的依存树包含了节点间的依存关系以及具体的依存关系类型. CTB 是国内外研究人员常用的一个大规模汉语树库,将其从短语结构转换为依存结构有利于人们进行汉语依存关系解析的研究. 本文在总结以往研究成果的基础上,结合 CTB 标注体系的特点,构造完整的汉语中心子节点过滤表,并基于中心子节点过滤表将 CTB 从短语结构转换为依存结构.

1 中心子节点过滤表

中心子节点在短语结构和依存结构中起着非常重要的作用. X-bar 理论[10]和管辖约束(GB)理论[11]等语言理论认为,每个短语结构中都有一个中心子节点决定着这个短语的主要性质,短语中的其他节点都是该中心子节点的修饰子节点. 而在依存结构中非中心子节点以某种依存关系依存于中心子节点.

依存结构树库 HIT-IR-CDT 标注体系如图 1 所示,其标注了句子中词语之间的依存关系及依存类型. 短语结构树库 CTB 中句子的标注如图 2 所示,其仅标出每个句子的短语层次结构及短语类型,没有标明每个短语的中心子节点. 确定短语中心子节点最常用的方法是使用中心子节点过滤表.

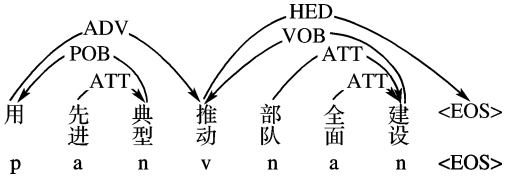

图 1 HIT-IR-CDT 依存结构树示例

Fig. 1 A dependency structure in the HIT Chinese Dependency Treebank

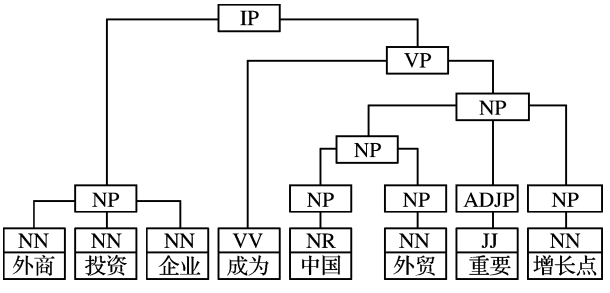

图 2 CTB 短语结构树示例

Fig. 2 A phrase structure in the Penn Chinese Treebank

1.1 中心子节点过滤表的结构

表 1 为部分中心子节点过滤表,其每一行都由〈短语类型,搜索方向,优先级〉三项组成. 其中,短语类型是非终端节点的短语符号;搜索方向为在非终端节点内部搜索中心子节点的方向,取值为 L 时从短语左侧开始向右搜索,取值为 R 时从短语右侧开始向左搜索;优先级为决定短语内部各类标注子节点作为中心节点的优先次序. 例如,根据过滤表中的一个条目〈IP:R, IP/VP/CP/VCD/VRD/VV/VA〉,可以这样确定 IP 短语的中心子节点:从右向左观察 IP 的每一个子节点,最先找到的标注为 IP 的子节点即为 IP 的中心子节点;如果没有找到 IP 节点,重新从右向左观察 IP 的每一个子节点,最先找到的符号为 VP 的子节点即为 IP 的中心子节点;以此类推,如果这个 IP 内部没有任何标注为 IP、VP、CP、VCD、VRD、VV、VA 的子节点,就默认最右侧的子节点为中心子节点.

表 1 中心子节点过滤表(示例)

Tab. 1 Head percolation table (examples)

短语类型	搜索方向	优先级
ADJP(形容词性短语)	R	ADJP/JJ
DNP("的"字短语)	R	DNP/DEG/DEV/DEC/NP
DP(限定词短语)	L	DP/CLP/QP/DT/DEG
INTJ	L	INTJ/IJ
IP(由 INFL 引导的从句)	R	IP/VP/CP/VCD/VRD/VV/VA
LCP(方位词短语)	R	LCP/LC/IP/NP/CLP
PP(介词短语)	L	PP/P/NP
QP(数量短语)	R	QP/CLP/CD/OD/DP/NP

1.2 中心子节点过滤表的制定

依据中文依存句法制定中心子节点过滤表. 本文采用了简单的依存关系描述体系,其中包含 10 种依存关系类型,如表 2 所示. 非中心子节点以表中所示依存关系类型依存于中心子节点.

表 2 举例说明了各类依存关系的定义. 其中"举例"一列中,标注为[1]的为例句 1"城乡游/首迎式/在/北京/举行/,市/领导/及/一千多/名/游客/出席/欢庆/活动。"中包含的依存关系. 标注为[2]的为例句 2"他/背起/沉重/的/行李/,匆匆/地/走出/家门。"中包含的依存关系. 句中分词依照 CTB 的标注体系. 箭头"→"左侧的词依存于其右侧的词.

列出 CTB 中每种短语包含的全部节点类型,依据上述依存关系的定义,对各类节点作为中心子节点的优先级进行排序,制定中心子节点过滤表,如表 1 所示.

表 2　依存关系类型

Tab. 2　The types of dependency relations

依存关系名称	依存关系标记	节点间依存与被依存关系	举例
主谓关系	SBV	主语依存于谓语	[1]游客→出席
动宾关系	VOB	宾语依存于动词	[1]活动→出席
定中关系	ATT	定语依存于中心词	[1]市→领导
介宾关系	POB	宾语依存于介词	[1]北京→在
连谓结构	VV	连谓的第一个动词依存于第二个动词	[2]背起→走出
状中结构	ADV	状语依存于中心词	[1]在→举行
数量关系	QUN	数词依存于量词	[1]一千多→名
并列关系	COO	并列关系的第一个词依存于第二个词	[1]领导→游客
“的”字结构	DE	依存于“的”	[2]沉重→的
“地”字结构	DI	依存于“地”	[2]匆匆→地

2　短语结构树到依存结构树的转换

制定了中心子节点过滤表，短语结构树到依存结构树的转换就相当直接. 转换采用递归算法，将短语结构树 Tree root 转换为依存结构树 Tree deprel 的转换算法 Convert(Tree root，Tree deprel)为

(1)若根节点 root 为叶子节点，返回根节点 root 并完成转换；

(2)查找根节点 root 的中心子节点；

(3)转换以中心子节点为根的子树，并返回该子树的中心子节点 headChild；

(4)对于其他非中心子节点

①转换以非中心子节点为根的子树，并返回该子树的中心子节点 non-headChild；

②将 non-headChild 依存于 headChild，并填入依存结构树 Tree deprel.

图 3 显示了图 2 所示例句从短语结构树到依存结构树的转换. 首先转换以 IP 为根节点的短语结构树 Tree IP. IP 不是叶子节点，查找 IP 短语的中心子节点. 参照表 1 中心子节点过滤表，从右向左观察‘IP’的每一个子节点，最先找到的标注为‘IP’的子节点即为‘IP’的中心子节点. 因为没有找到‘IP’子节点，重新从右向左查找标注为‘VP’的子节点为‘IP’的中心子节点(图中实线所示). 然后转换中心子树 Tree VP，‘VP’的中心子节点‘VV(成为)’为叶子节点，将其返回. 继续转换‘VP’的非中心子树 Tree NP… 依次确定句子每个短语的中心子节点，将短语的非中心子节点依存到其中心子节点上. 如短语‘NP(NP(NR(中国))NP(NN(外贸)))’的中心子节点为‘NN(外贸)’，则节点‘NN(外贸)’决定了短语‘NP(NP(NR(中国))NP(NN(外贸)))’的主要性质，非中心子节点‘NR(中国)’依存于中心子节点‘NN(外贸)’. 其上一层的短语‘NP(NP(NR(中国))NP(NN(外贸)))ADJP(JJ(重要))NP(NN(增长点))’的中心子节点为最右侧的短语‘NP(NN(增长点))’，非中心子节点‘JJ(重要)’依存于中心子节点‘NN(增长点)’；同理，另外一个非中心子节点‘NP(NR(中国)NN(外贸))’的中心子节点‘NN(外贸)’也依存于‘NN(增长点)’. 最后得到整句的依存关系：“NN(外商)→NN(企业)；NN(投资)→NN(企业)；NR(中国)→NN(外贸)；NN(外贸)→NN(增长点)；JJ(重要)→NN(增长点)；NN(增长点)→VP(成为)；NN(企业)→VV(成为)”. 节点‘VV(成为)’为整句的中心子节点.

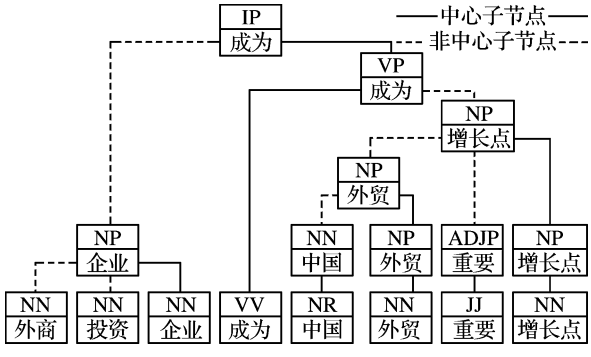

图 3　短语结构树到依存结构树的转换例句

Fig. 3　An example of converting phrase structure tree to dependency structure tree

3　实　验

本文对宾州中文树库(CTB5.0)进行了转换. CTB5.0 是在美国国防部资助下，由语言数据联盟(Linguistic Data Consortium，LDC)发布的一个语料库. 其由 890 个文件组成，包含 18 782 个句子，507 222 个词，824 983 个汉字. 句子平均长

度是 27 个词. 句子如图 2 所示,标注了分词、词性和短语结构. 文章大部分来自《新华日报》,还有少部分来自香港地区政府和台湾地区媒体. 文章覆盖政治、经济和文化等领域. 本文将 CTB 中的短语按类型进行统计,统计结果见表 3. 因为短语 WHNP 和 WHPP 没有子节点,所以不包含在表 1 中.

表 3 CTB 短语类型及数量统计

Tab. 3 Details of the phrases in the Penn Chinese Treebank

短语类型	数量	短语类型	数量	短语类型	数量	短语类型	数量
ADJP	13 590	INTJ	13	PRN	1 182	VPT	294
ADVP	37 203	IP	75 508	QP	20 217	VRD	1 360
CLP	14 380	LCP	7 983	UCP	375	VSB	450
CP	25 370	LST	211	VCD	767	FRAG	1 963
DNP	12 509	MSP	1 343	VCP	70	WHNP	11 592
DP	6 127	NP	234 183	VNV	127	WHPP	339
DVP	707	PP	18 699	VP	134 871		

为检验转换效果,随机地从 CTB5.0 中取出 200 句语料进行测试,句子的平均句长同样是 27 个词. 对测试语料中出现的短语的转换结果进行了人工校验,转换正确率如表 4 所示. 部分短语类型在测试语料中没有出现.

由表 4 可见,大部分短语的转换正确率接近或达到了 100%,测试语料中所有短语的转换正确率达到 99.50%. CP、IP、NP 节点转换依然存在错误,其他节点包括大量的 VP 节点,正确率均为 100%.

表 4 转换正确率

Tab. 4 Conversion precision

短语类型	样本数目	正确数目	正确率/%	短语类型	样本数目	正确数目	正确率/%
ADJP	205	205	100	NP	2 351	2 348	99.87
ADVP	265	265	100	PP	213	213	100
CLP	228	228	100	PRN	10	10	100
CP	234	233	99.57	QP	301	301	100
DNP	108	108	100	UCP	5	5	100
DP	83	83	100	VCD	11	11	100
DVP	3	3	100	VP	1 121	1 121	100
IP	664	662	99.70	VRD	16	16	100
LCP	72	72	100	VSB	6	6	100

通过对转换错误的短语结构进行观察发现,大部分是处于外层的具有复杂结构的短语. 如图 4 所示 IP 短语包含两个 IP 子节点. 依据中心子节点过滤表,右侧‘IP’子节点为外层 IP 短语的中心子节点,左侧‘IP’子节点的中心节点‘VV(达)’依存于右侧‘IP’子节点的中心节点‘M(美元)’. 但是这种依存关系违背了依存文法. 造成这种错误的原因是右侧‘IP’短语没有动词,所以‘M(美元)’成为其中心节点,使得‘VV(达)’错误地依存于‘M(美元)’. 如果右侧‘IP’子节点添加了动词,改为“投资额/约/达/一千万/美元”,则其中的‘VV(达)’将成为其中心节点. 最终,左侧‘IP’的中心节点‘VV(达)’依存于右侧‘IP’的中心节点‘VV(达)’,这种依存关系符合依存文法.

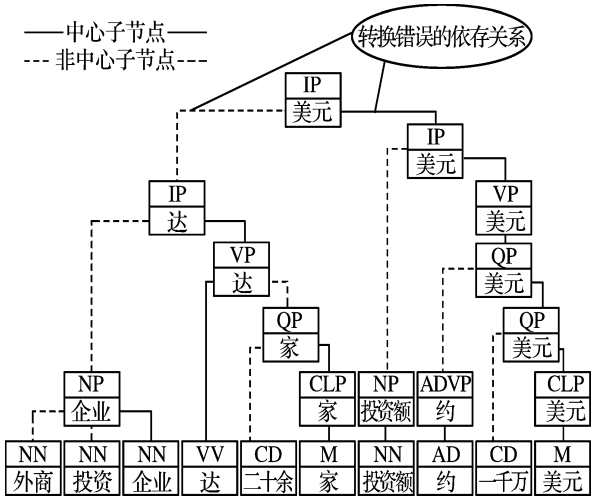

图 4 转换错误例句

Fig. 4 An example of the conversion error

Xia 基于中心子节点过滤表将 PTB 和 CTB 从短语结构树转换为依存结构树[8]. 其指出除了具有同位语的名词短语(NP)外，PTB 中其他短语都取得了较好的转换效果，但是没有给出具体的转换正确率. Xia 没有提及 CTB 的转换效果.

4　结论和展望

本文介绍了将 CTB 从短语结构转换为依存结构的方法. 基于现代中文依存关系文法制定了完整的中心子节点过滤表，依据该表可以确定 CTB 中所有短语的中心子节点，并将短语结构的 CTB 转换为依存结构. 抽样测试结果表明转换正确率达到了 99.50%. 本文的工作，使得短语结构的 CTB 可以用于中文依存关系解析的研究，对于研究句法解析具有一定的参考价值. 下一步，将利用转换得到的依存结构树库，进行中文依存关系解析的研究.

参考文献：

[1] 周　强，赵颖泽. 汉语功能块自动分析 [J]. 中文信息学报，2007，**21**(5)：18-24

[2] CHENG Yu-chang，ASAHARA M，MATSUMOTO Y. Machine learning-based dependency analyzer for Chinese [C] // MINGHUI D，HAIZHOU L，MIN Z，eds. **Proceedings of the International Conference on Chinese Computing 2005**. Singapore：COLIPS Publication，2005：66-73

[3] 周惠巍，杨　洋，黄德根. 基于远距离依存关系的中文依存关系解析[J]. 计算机工程，2007，**33**(24)：212-214

[4] XUE Nian-wen，XIA Fei，CHIOU Fu-dong，*et al*. The Penn Chinese Treebank：phrase structure annotation of a large corpus [J]. **Natural Language Engineering**，2005，**11**(2)：207-238

[5] 曹海龙，赵铁军，李　生. 基于中心驱动模型的宾州中文树库(CTB) 句法分析 [J]. 高技术通讯，2007，**17**(1)：15-20

[6] CHENG Yu-chang，ASAHARA M，MATSUMOTO Y. Chinese deterministic dependency analyzer：examining effects of global features and root node finder [C] // **Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing**. Korea：SIGHAN，2005：17-24

[7] LIN De-kang. A dependency-based method for evaluating broad-coverage parsers [J]. **Natural Language Engineering**，1998，**4**(2)：97-114

[8] XIA Fei. Automatic grammar generation from two different perspectives [D]. Philadelphia：University of Pennsylvania，1999

[9] 党政法，周　强. 短语树到依存树的自动转换研究 [J]. 中文信息学报，2005，**19**(3)：21-27

[10] CHOMSKY N. Remarks on nominalization [C] // JACOBS R，ROSENBAUM P，eds. **Reading in English Transformational Grammar**. Waltham (MA)：Ginn and Co.，1970：184-221

[11] CHOMSKY N. **Lectures on Government and Binding** [M]. Dordrecht：Foris Publications，1981

Research on treebank converting from phrase structures to dependency structures

ZHOU Hui-wei*，HUANG De-gen，QIAN Zhi-qiang，YANG Yuan-sheng

(School of Computer Science and Technology，Dalian University of Technology，Dalian 116024，China)

Abstract：Large-scale phrase structure treebank and dependency structure treebank are developed and interconverted for the purpose of syntactic analysis on true corpus. The head percolation table is constructed on modern Chinese dependency grammar by discussing the relationship between phrase structure and dependency structure based on Penn Chinese Treebank (CTB)，and CTB from phrase structure is converted to dependency structure treebank using the head percolation table. 200 sentences are chosen from CTB randomly to evaluate the conversion performance. Precision of the conversion has attained 99.50%. The achieved dependency structure treebank can be used to analyze Chinese dependency relation.

Key words：syntactic parsing；phrase structure；dependency structure；treebank；converting