短语结构树到依存树的转换

王跃龙　韩　希

（武汉大学语言与信息研究中心，武汉　430072）

摘　要： 短语结构树与依存树之间的转换是计算语言学重要的研究内容之一。依存树可能是表示汉语句法结构更好的形式。本文尝试把宾州汉语树库(4.0)的短语结构树转换为依存树。介绍了具体的操作步骤、转换规则和算法，指出了预处理的重要性，并在传统的依存表示上做出了部分的修正，保留了部分非终极节点，使之更适于向语义关系的转换，依存关系采用带功能标记的三个非终极符号表示。转换得到的依存树可以有不同的颗粒度。整个依存树库具有很强的适用性。

关键词： 树库；短语结构树；依存树；自动转换；

Transforming phrase-structure trees into dependency trees

Wang Yuelong　Han Xi

(Center for the study of language and information Wuhan University, Wuhan 430072)

Abstract: Converting between phrase-structure tree and dependency tree is one of the most important subjects in computational linguistics. Dependency tree maybe is a better form to express Chinese. This paper tried to transform U-penn Chinese Treebank to dependency treebank. This paper introduced the procedure, rules and algorithms . This paper pointed the importance of preprocessing, and revised former dependency theory , kept some non-terminal nodes. Transformed dependency treebank is fit for transforming into semantic level. Dependency relationships were expressed by three non-terminal signs with function marks, which can adjust according to different need to obtain the objectivity of measure.

keywords: Treebank ; phrase-structure Treebank ; dependency tree; automatically convert;

1．引言

树库的表示方法主要有两种，一种是短语结构树，一种是依存树。在世界范围来说，大多数大规模树库是基于短语结构的。在关于汉语的树库中，基于短语结构标注的树库也占据着主要的地位。

短语结构树库并不是比较和评价句法分析器的最理想形式[1]，依存结构树库允许更有意义的测评和比较。依存的结构表示，易于有选择的评价句法分析器的表现。还可以避免掉一些无意义的分歧，例如对在某一特定语言中是否存在动词短语结构的争论等。

作者简介：王跃龙(1979--)男,河北石家庄人，硕士研究生，E-mail:wangyuelong_2001@126.com

韩　希(1983--)男,湖北鄂州人，　硕士研究生，E-mail:hanxi2008@163.com

认知语言学研究表明，依存语法是跟人的认知习惯相扣合的。依存树相对于短语结构树来说有许多自己的优势，依存树没有非终极节点，可以节省大量的存储空间，依存树注重的是外部的关系，短语结构注重的是内部的结构。依存结构的表示与语义的表示之间有一种比较天然的对应关系。有利于向格关系的转换等。依存树库的构建也是自然语言处理应用方面的需要。机器翻译和问答系统更多的是要理解句子的谓词论元结构。所以依存树库的构建是非常有必要的。

2．前人的转换实践

短语结构的分析是一种自上而下的分析方法，依存分析法则是自下而上的分析方法。自 Gaifman 指出短语结构跟依存结构之间存在着等同关系以后，很多学者都对转换提出了不同的算法。

比较早的有 Lin 的将短语结构树库转换为依存结构树库的算法，但是他的方法不是完整的算法，对寻找中心子节点的方法没有提及。xia fei [2]采用中心词过滤表的方法完成了将短语结构树库转换为依存树库，但是没有表示出两个节点之间具体的依存关系类型。

Michael Daum et al. 在应用工具 DEPSY 把 NEGRA 树库转换为依存结构时则是运用到了转换规则表，并运用依存关系表来决定依存关系的标注[3]。Žabokrtský, Z.和 Kučerová, I [4]把宾州英语树库的一部分转换为类似于布拉格依存树库（Prague Dependency Treebank）标注规范的依存树。

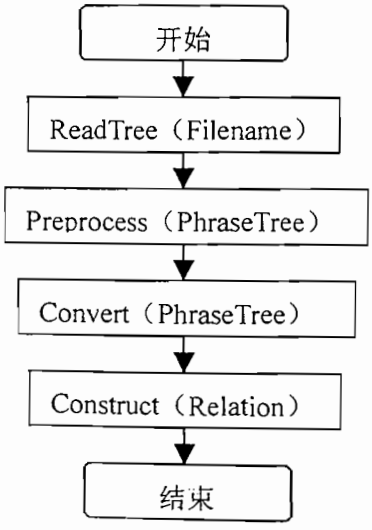
早期的转换大部分都是针对英语语料库的。Eisner（1996）运用从宾州树库（英语）中转换得到的依存结构来训练相对应的随机句法分析器。Collins (1999)根据 Lin 的最初提议，把金标准（gold standard）和句法分析结果都转换为依存结构。后来此方法也用来评价基于 TIGER 树库的部分句法分析器(Kubler and Telljohann,2002)。后来别的语言的树库转换才逐渐多起来。针对汉语树库的转换仅见于把 TCT 转换为依存的树库，其依存关系的确定依据其原来短语结构树的结构标记[5]。

3. 短语结构树到依存树的转换

我们尝试把宾州汉语树库 4.0 版的短语结构树转换为依存结构树。转换的步骤：

1．把短语结构表示为直观的树形图；
2．提取标点，提取基本名词短语，只保留基本名词短语的中心词，并且对一些特殊的语法现象进行预处理；
3．根据规则找出每一个非终极节点的中心子节点；
4．根据算法找出整理后的依存树中词语之间的依存关系。
5．根据依存关系形成依存树。
6．分析基本名词短语，表示为依存树；
7．把基本名词短语的依存子树对接于句子的依存树上，形成总的依存树。

整个转换过程如以流程图图 1 所示：



ReadTree（Filename）：读取宾州汉语树库文件，将树库中的短语结构树读入，用一个结构体来表示各结点的信息，然后按树库中树的形式把结点连接起来构建成便于计算机处理的树，对于树库中表示省略的空节点，也用相同的结构体保存并挂在新生成的短语结构树上。树中的叶子结点，即终极结点为句子中的实际的词。

```
struct node{
    CString type;              //表示该结点的短语结构标注信息
    CString mword;             //该结点的中心词
    node *llink, *rlink,*link;  //该结点的链接指针，用于构建树
};
```

Preprocess（PhraseTree）：预处理包括：

1 抽取掉标点，从短语结构树去掉 type 为 PU 的节点；

2 抽取基本名词短语，直接用中心词（默认取名词短语中最后一个名词）替换基本名词短语并列所在的分枝，然后对该分枝各成分进行分析；

3 “把”字句和“被”字句的处理，宾州树库中“把”字句和长“被”字句分别标注为"BA"和"LB"，"BA"结构一般出现在 VP<BA IP-obj>中，此时将"BA"从树上拿下，作为"IP-obj"的子结点挂在"IP-obj"结点下；

4 把主句和分句分开，搜寻分句中的关联词；

5 提取句末的语气词，用于最后依存在依存树的根节点上；

6 特殊语法情况，保留非终极结点，比如句子做主语（IP -SBJ）或宾语，并列结构（如 VCD）等；

7 保留 VNV 、VPT，作为整体处理。

4. 转换的规则举例

表 1 转换的对应规则（举例）

Tab.1 Transformational rules

	转换规则		依存关系
PP		PP<ADVP !PP>	<ADVP PP !PP>
	PP<!P XP>	PP<!P NP>	<!P PP NP>
		PP<!P LCP>	<!P PP LCP>
		PP<!P IP>	<!P PP IP>

5. 转换的几点解释

5.1 转换之前的预处理

转换之前的预处理非常重要。预处理包括抽取掉标点，把主句和分句分来，抽取基本名词短语，对一些特殊语法现象作预处理等。宾州汉语树库最初构建是根据英语树库的框架来的，所以“把许多汉语独具特色的描述信息硬纳入英语的描述框架，总给以汉语为母语的人许多生硬的感觉”[6]。为使宾州汉语树库能够符合我们的分析习惯，必须在转换前做出必需的预处理。

5.2 部分非终极节点在依存树结构中的保留问题

鉴于依存的表示是为了更方便地向语义关系转换，我们在转换时就应该尽量考虑符合直观的语义关系。我们在特定的语言现象中引入非终极节点。这样的表示也有其心理学的根源和依据。典型的如：句子做主语或宾语，我们就通过保留非终极节点来表示这个句子。我们在并列结构的处理中，也保留了非终极节点，另保留 VNV ，作为整体处理。另外我们还保留了 FRAG 结构。

5.3 保留了表示空语类的空节点

宾州汉语树库中有表示空语类的空节点，在转换为依存树以后，我们保留了这些空节点，这样有利于直观的观察动词的配价情况以及词语之间的同指。

5.4 宾州汉语树库与其他汉语树库处理不同之处以及应对策略

在我们转换的依存树中采用多重依存，允许兼语成分同时依存于两个动词，即是第一个动词的宾语又是第二个动词的主语。宾州树库的处理中把一个句子的分句当作主句的修饰成分，在转换为时，我们把它表示为两棵树，

再标注两颗树之间的关系。

把字结构和被字结构中，宾州树库把引介的宾语和后边的动词一起作为一个句子来处理。我们把把字句在转换之前作豫处理。最后得到的依存树中中心词是动词，“把”依存于中心动词，“把”的宾语依存于“把”。另外，一些句末的语气词，在转换之前先提取，最后依存在树的根节点上。

5.5 用三个非终极符号来表示之间的依存关系

为了增加我们转换后得到的依存树库和不同的语法体系不同表示方式的兼容性，我们采用三个非终极符号来表示两个词语之间的依存关系，这样具有一定的弹性，可以根据具体情况作出调整，在运用于对一些句法分析器的评测时会有更强的客观性。

6. 转换错误的分析

转换后的错误多跟原来短语结构树的标注有关。单纯由于算法导致的错误几乎没有发现。通过转换可以考察到原来树库标注的不足，进而做出改进。例如：把关联词语与句子放在同一个层面，转换后表示分句间的关系就可能出错。再一个就是基本名词短语的处理，例如：(NP (NN 开发)(CC 与) (NN 法制) (NN 建设)) 这个名词短语转换为依存结构为图 2 是错误的，正确的转换应该是图 3。这个错误的发生就是跟原始语料的标注有密切的关系，原来短语结构树的标注颗粒度较粗。

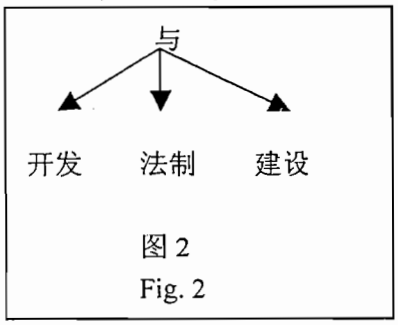


图 2
Fig. 2

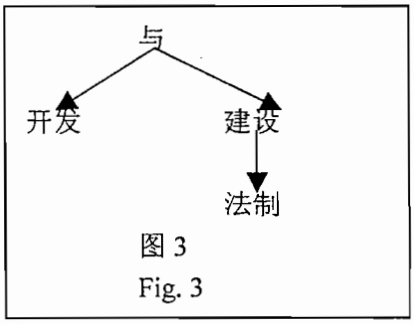


图 3
Fig. 3

7. 结语

把宾州汉语树库 4.0 转换为依存树库，可以使宾州树库在更广阔的范围内得以应用，并且发现其原来标注体系的不足，使其不断改进。转换后的依存树库，将更便于满足应用的需求。在今后的工作中，我们也会不断改进我们的转换方法，使其更接近我们所构想的理想状态依存结构树。在下一步的工作中，我们将尝试进行浅层的语义转换，从句法层面深入到语义层面。

参考文献

[1] Lin, D. A dependency based method for evaluating broad-coverage parsers [A].In: Proceedings of IJCAI95[C], Montreal , Quebec, Canada,1995

[2]Fei Xia and Martha Palmer . Converting Dependency Structures to Phrase Structures[A]. In proceedings of the Human Language Technology Conference (HLT-2001),San Diego,CA,2001,March 18--21

[3] Daum, M, Foth, K. and Menzel, W . Automatic transformation of phrase treebanks to dependency trees[A]. In Proc. 4th Int. Conference on Language Resources and Evaluation (LREC 2004).

[4]Žabokrtský, Z; Kučerová, I: Transforming Penn Treebank Phrase Trees into (Praguian) Tectogrammatical Dependency Trees[A]. PBML 2002 (Prague Bulletin of Mathematical Linguistics), 2002

[5] 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)；

[6] 周强，汉语句法树库标注体系[J]，中文信息学报，2004,18(4)；