文章编号：1003-0077(2023)02-0015-11

句式结构树库的自动构建研究

谢晨晖[1,2,3]，胡正升[1,2,3]，杨麟儿[1,2,3]，廖田昕[1,2,3]，杨尔弘[1,3]

(1. 北京语言大学 国家语言资源监测与研究平面媒体中心，北京 100083；
2. 北京语言大学 信息科学学院，北京 100083；
3. 北京语言大学 语言资源高精尖创新中心，北京 100083)

摘 要：句式结构树库是以句本位语法为理论基础构建的句法资源，对汉语教学以及句式结构自动句法分析等研究具有重要意义。目前已有的句式结构树库语料主要来源于教材领域，其他领域的标注数据较为缺乏，如何高效地扩充高质量的句法树库是值得研究的问题。人工标注句法树库费时费力，树库质量也难以保证，为此，该文尝试通过规则的方法，将宾州中文树库 (CTB) 转换为句式结构树库，从而扩大现有句式结构树库的规模。实验结果表明，该文提出的基于树库转换规则的方法是有效的。

关键词：句式结构；短语结构；树库构建

中图分类号：TP391 **文献标识码**：A

Automatic Construction of Sentence Pattern Structure Treebank

XIE Chenhui[1,2,3], HU Zhengsheng[1,2,3], YANG Lin'er[1,2,3], LIAO Tianxin[1,2,3], YANG Erhong[1,3]

(1. National Language Resources Monitoring and Research Center Print Media Language Branch, Beijing Language and Culture University, Beijing 100083, China;
2. School of Information Science, Beijing Language and Culture University, Beijing 100083, China;
3. Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China)

Abstract: Sentence pattern structure treebank is developed according to the theory of sentence-based grammar, which is of great significance to Chinese teaching. To further expand such treebank from Chinese as second language textbooks and Chinese textbooks to other domains, we propose a rule-based method to convert a phrase structure treebank named Penn Chinese Treebank (CTB) into a sentence pattern structure treebank so as to increase the size of the existing treebank. The experimental results show that our proposed method is effective.

Keywords: sentence pattern structure; phrase structure; treebank construction

0 引言

树库是经过标注的深加工语料库，其记录着真实文本中每个句子的句法标注结果，提供分词、词性标注、句法结构等信息。目前广泛应用于自然语言处理领域的主流树库类型是短语结构树库和依存结构树库，句式结构树库的资源较少。

短语结构树库遵循短语结构语法，描写句子的短语结构，中文领域具有代表性的是宾州中文树库(Penn Chinese Treebank, CTB)[1]、清华汉语树库(Tsinghua Chinese Treebank, TCT)[2]等；依存结构树库遵循依存语法理论，该理论由法国语言学家L.Tesnière于1959年提出，通过句子中词与词之间的依存关系来分析句法结构，认为任何两个词之间的依存关系中必有一个是中心词[3]，具有代表性的

收稿日期：2022-11-16 **定稿日期**：2022-12-16

基金项目：国家语委项目(ZDI135-131)；中外语言交流合作中心《国际中文教育中文水平等级标准》教学资源建设项目(YHJC21YB-129)；中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)；北京语言大学研究生创新基金(中央高校基本科研业务费专项资金)项目成果(22YCX086)

中文依存树库是哈尔滨工业大学汉语依存树库(HIT-IR-CDT, CDT)[4]。

值得注意的是,短语结构语法和依存结构语法均"以结构关系描写代替句子格局分析","句式结构在中文信息处理中一直处于一种模糊的地位"[5],鉴于此,北京师范大学语言与文字资源研究中心构建了句式结构树库,弥补了这一不足。该树库以句本位语法体系为理论基础,着力研究各类句式的结构规律,总体特点为:①句法上以句子成分分析法作为析句方法,并以"图解法"为语法分析工具;②词法上采用"依句辨品"的词法观,"以句法控制词法",这种语法本质上属于教学语法。

20世纪60年代,受到结构主义描写语言学的影响,汉语的句法分析开始以"直接成分"分析取代"句子成分"分析,典型的是朱德熙先生在《语法讲义》中建立的"词组本位"语法体系[6],其以"短语"为本位作为汉语语法分析的基础,总体特点为:①句法上以直接成分分析法(或称层次分析法)为析句方法,认为汉语句子构造原则与词组构造原则一致;②词法上,以词的语法功能为划分词类的标准。

直接成分分析法属于结构主义语言学的句法分析方法,短语结构语法实际上也是从直接成分分析法派生而来。整体而言,与短语结构树库标注句子层次结构、依存结构树库描写词与词之间关系相比,句式结构树库更能够呈现句子的整体结构,树结构更加扁平。

大规模树库多以自动句法分析为主要应用,但对于句式结构树库,"语言教学既是其理论之源,也是主要应用方向之一"[7],如交互式出题、文本可读性评估等研究,句式结构自动分析器的研究还处于初步阶段。目前该树库语料约7万句,主要来源于国际汉语教材、中小学语文教材、文学作品等,其他领域语料较为缺乏,如何高效扩充出高质量的句法树库是值得探究的问题。

大规模句法资源的构建是一项费时费力的工程,目前常用的有人工标注和树库转换两种方法。人工标注树库能够保证树库质量,但成本高,耗时长。第二种方法即利用现有的树库资源,通过寻找两种形式语法之间的映射关系,转化成所需的目标树库,这种方法更加高效。理论上来说,不同类型的树库尽管在语法形式上各不相同,但本质上都是对真实文本的句法结构的描写,这使得不同树库的转换具有可行性。

目前关于树库的自动转换研究主要集中在短语树库与依存树库之间的转换。Lin[8]在1995年较早地提出了一种中心词节点表的方法,将短语树转换到依存树;Xia等人[9]阐述了两种将短语树转成依存树的算法,采用中心词过滤表的方法将宾州树库(Penn Treebank, PTB)转换成依存树库,并提出一种新的算法,将产生的依存树转换成短语树,转换结果很接近原有的PTB。此外,Žabokrtský等人[10]、Niu等人[11]、以及Kong等人[12]对短语结构与依存结构之间转换也做了研究与实践。在中文领域,树库转换研究较早的是党政法等人[13],其在Lin[8]与Xia[9]的研究基础上结合TCT的特点,进一步完善了转换算法,将TCT转换成了依存结构,转换准确率达97.37%;李正华等人[14]通过统计与规则相结合的方法,将CTB转换成哈工大依存树库体系结构;周惠巍等人[15]在Xia等人[9]提出的中心词过滤表方法及前人研究的基础上,结合CTB的特点,构造了完整的汉语中心词过滤表,将CTB转成了依存结构树库。

相比之下,短语结构树库、依存结构树库同句式结构树库之间的转换研究较少,其中张引兵等人[16]通过总结TCT与句式结构树库标注体系的映射关系,制定了一套转换规则,将TCT转换成了句式结构树库,总体准确率为92.9%,验证了短语结构向句式结构转换的可行性。

考虑到宾州中文树库(CTB)在自然语言处理领域的通用性,本文通过比较宾州中文树库与句式结构树库在语法形式上的异同,制定了树库的自动转换规则,将短语结构树库CTB自动转换为了句式结构树库,实现了新闻领域句式结构树库的自动构建。为了验证树库转换规则的有效性,本文在人工标注的测试集上进行了三组实验,以比较基于句式结构自动句法分析的方法、基于短语结构自动句法分析结合转换规则的方法和基于宾州中文树库结合转换规则的方法三者的效果。这三种方法在测试集上的F_1值分别为84.43%、87.56%、89.72%,说明本文提出的基于转换规则的方法是有效的①。

1 背景介绍

本文所使用的源树库为宾州中文树库,是短语结构树库的代表。制定向句式结构树库的转换

① 本文代码已公开在GitHub平台上,网址为:https://github.com/blcuicall/ctb2stb

规则需要比较两者在语法形式表现上的异同，短语结构树库与句式结构树库的最本质区别在于两者所遵循的语法体系的不同。短语结构树库遵循短语结构语法，这是乔姆斯基为说明转换生成语法而讨论的一种语法模式[17]，其以结构主义语言学的“直接成分分析法”为基础，以更接近数学公式的重写规则表示短语以及句子的结构。句式结构树库遵循句本位语法理论，使用“句子成分分析法”来分析句子，归纳句型句式，这种析句方法属于汉语传统语法。具体而言，宾州中文树库与句式结构树库在标注体系、句法分析单位、句法关系的描写等方面存在差异。

1.1　宾州中文树库

宾州中文树库是美国宾州大学自 1998 年起构建的短语结构树库，简称 CTB，该树库以短语为句法分析单位，理论基础为语杠理论（X-bar Theory）和支配及约束理论（Government and Binding Theory）[1]，但作了一些简约化处理，不完全遵循二分法。CTB 包含新华社、政府文件、新闻杂志、广播、访谈、网络新闻及网络日志等内容，对短语结构、短语功能进行了详细标注。

宾州中文树库通过括号的嵌套来存储层级结构。从图 1 可以看出，除了词性节点以外，非叶子节点上的典型标记格式为“短语标签-功能标签”，树库中还存在“短语标记-多个功能标记”组合的情况。CTB 注重短语外部功能的描写，例如，“NP-SBJ”，“NP”表示该节点为名词性短语，“SBJ”表示它与其右侧兄弟节点 VP 之间为主谓关系，但其父节点却并没有相应的标记来指明这层关系，而是选择将其蕴含于单个功能标签当中，如主谓关系蕴含于“SBJ”中，述宾关系蕴含于“OBJ”中等。直观来看，部分节点上的功能标记可以直接对应汉语传统语法中的句子成分（如主语和宾语）。

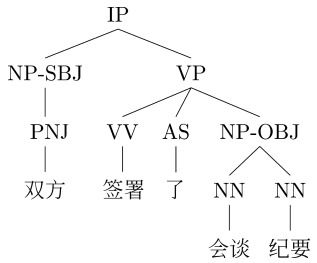

图 1　短语结构树库

相比于句式结构树库，短语结构树库更侧重于描写短语结构，在句式信息的整体表现上不可避免地有所缺失，如“兼语句”“连动句”等特殊句式无可利用的标记，只能通过句法结构层次呈现。

1.2　句式结构树库

句式结构树库由北京师范大学语言与文字资源研究中心构建，该树库遵循句本位语法理论，因此也称“句本位语法树库”。句本位语法由黎锦熙先生在 1924 年出版的《新著国语文法》中提出，该理论主要面向语法教学，以“句子成分分析法”为析句方法，主张“先理会综合的宏纲（句子），再从事于分析的细目（词类）”[18]，并以图解法作为析句工具。句式结构树库语料主要来自具有一定影响力的国际汉语教材，也包括中小学语文教材，目前规模约 7 万句。相较于短语结构树库，句式结构树库以六大句法成分为基础构建句子结构，在词法层面，句式结构树库区分词库词与动态词，并对动态词内部词法结构进行了详细标注。

句式结构树库以 XML 格式进行存储，本质上也是树结构。以图 2 为例，树的根节点为＜ju＞，表示一句话的开始，当句子有多个小句时，首先将其拆分为各小句＜xj＞，然后对每个小句进行分析，该例只有一个小句。句本位语法的特征之一是采用“图解法”表示语法分析结果，如图 3 所示。其中，双竖线左侧为主语，单竖线右侧为宾语，横线上方为句子的主干结构。

```
<ju txt="双方签署了会谈纪要">
  <xj>
    <sbj>
      <r>双方</r>
    </sbj>
    <prd scp="VO">
      <v>签署</v>
    </prd>
    <uv>
      <u>了</u>
    </uv>
    <obj>
      <n>会谈</n>
      <n>纪要</n>
    </obj>
  </xj>
</ju>
```

图 2　句式结构树库

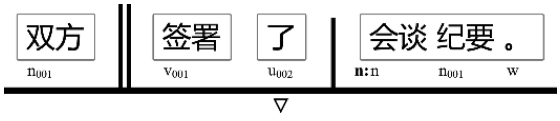

图 3　句式结构图解

句式结构树库注重对“句式”的描写，根据附加成分的有无、谓核的数量，将句子分为“基本句式”“扩展句式”和“复杂句式”。基本句式为简单的主谓宾结构(图3)；扩展句式在此基础上增加附加成分——定语、状语或补语。＜xj＞下的最外层以主＜sbj＞、谓＜prd＞、宾＜obj＞为句子主干成分，以定＜att＞、状＜adv＞、补＜cmp＞为附加成分，同时为介词(p)、连词(c)、助词(u)、方位词(f)设置了“虚词位”，如图中的助词位＜uv＞＜u＞了＜/u＞＜/uv＞。以上两种句式都是“单谓核句”，只有一个谓核，如图3的“签署”。复杂句式指多谓谓语句、主谓谓语句和复句，包含多个谓核，并根据谓核的关系定义细类，如连动句、兼语句等。在标记上，树库通过属性标记“scp”和“fun”保存句式信息，如图2中谓核“签署”的属性标记“scp”取值为“VO”，说明该谓核后接宾语。当句子中有多个谓核时，谓核之间需要通过属性标记“fun”表明其续接关系。

2 构建句式结构树库

我们采用宾州中文树库(CTB)作为源树库。将源树库转为目标树库总体上需要包含两部分：词层面的转换和句法层面的转换。但CTB与句式结构在词法层面的差别较大，包括分词粒度与词语结构标注的差异。具体来说，句式结构树库区分词库词与动态词，并标注动态词的内部结构，CTB对词法结构的标注却很模糊，基本上只显示词语边界而不含词语结构，如复合词呈现出扁平的树结构。因此，词法层面的转换需要实现分词粒度的统一与动态词的识别。考虑到转换工作难度与工作量，我们采取先句法转换后词法转换的策略：①在句法上完全依照句式结构树库的规范；②由于分词粒度的改变可能涉及句法标注的改变，为了尽可能利用现有的CTB的句法标注信息，词法上暂时保留源树库CTB的分词与词性特征。下面将详细介绍我们制定的自动转换规则及转换算法。

2.1 句法成分的转换规则

与句式结构树库以句子成分作为树的非叶子节点不同，CTB的句子成分信息分散在各类标记中，部分句式信息则通过括号的层级关系体现，因此需要分情况制定互补的转换规则。对于词性标记以外的非叶子节点的转换，从两个方面进行：句法成分信息和句式信息。在句法成分的提取上，利用CTB的功能标签集和短语标签集作为核心转换方法，利用CTB的词性标签集作为辅助转换方法，以补充核心转换方法的不足；在句式信息的提取上，主要利用各类句式的特殊标记或相对固定的树层级结构。从宾州中文树库向句式结构树库转换的总体思路是：将复句切分为小句，对CTB的短语树进行先序遍历，对于词性标记以外的非叶子节点，提取标签，对应转换规则进行转换；对于叶子节点(包括标点)，根据词类对应关系转换，即将“(词性 词)”的括号形式转换为＜词性＞词＜/词性＞的XML形式。我们根据所利用的标签类型分为核心转换规则和辅助转换规则。

2.1.1 核心转换规则

利用CTB功能标签和短语标签进行转换的规则为核心转换规则。在CTB的功能标签集里，有一部分如主语(SBJ)、宾语(OBJ)、状语(ADV)标签，可以与句法成分直接对应，这类功能标签可以直接转换为相应的句法成分；在短语标签集里，也有一部分用于指明短语的句法性质和内部的结构关系，例如DVP(“地”字短语)，如果DVP后跟有兄弟节点VP(动词短语)，则DVP为状语。类似的还有ADVP(副词短语)作状语，PP(介词短语)作状语或补语等，具体规则详见表1。

2.1.2 辅助转换规则

辅助转换规则指利用CTB的词性标签，将标签下的叶子节点转换为句本位语法中的句法成分，具体规则见表2。

句式结构树库为介、连、助、方位词设置虚词位：介词位、连词位、助词位、方位词位。虚词不直接充当句子成分，但在句式结构树库中的占位与主语等句子成分一致，因此我们将它纳入句子成分的规则集中。由于虚词位只需要获得词性信息，所以只要根据词性的对应关系即可进行简单转换。

表 1　核心转换规则

句式结构标记	含义	功能标记	短语标记	转换规则	举例
＜sbj＞	主语	SBJ TPC	—	＜sbj＞【XP-SBJ】＜/sbj＞ ＜sbj＞【XP-TPC】＜/sbj＞	【总理】说 【中国】发展迅速
＜obj＞	宾语	OBJ IO PRD	—	＜obj＞【XP-OBJ】＜/obj＞ ＜obj＞【XP-IO】＜/obj＞ ＜obj＞【XP-SBJ】＜/obj＞	抵达【罗马】 给【他】一本书 第二步是【集训】
＜att＞	定语	—	DNP QP DP	＜att＞【DNP】＜/att＞ ＜att＞【QP】＜/att＞ ＜att＞【DP】＜/att＞	【两国的】关系 【一个】食堂 【这次】大海啸
＜adv＞	状语	ADV TMP BNF DIR MNR	DVP ADVP	＜adv＞【XP-ADV】＜/adv＞ ＜adv＞【XP-TMP】＜/adv＞ ＜adv＞【XP-BNF】＜/adv＞ ＜adv＞【XP-DIR】＜/adv＞ ＜adv＞【XP-MNR】＜/adv＞ ＜adv＞【DVP】＜/adv＞ ＜adv＞【ADVP】＜/adv＞	【跟去年】相比的话 【8 月 10 日】宣称 【给灾民】送礼物 【往山上】跑 【以各种方式】参与 【热切地】盼望 【立刻】跳下车

注：XP 指 CTB 中的任何短语标记，下同。

表 2　辅助转换规则

标记	含义	词性标记	转换规则	举例
＜prd＞	谓语	VC VE VV	＜prd＞＜v＞【VC】＜/v＞＜/prd＞ ＜prd＞＜v＞【VE】＜/v＞＜/prd＞ ＜prd＞＜v＞【VV】＜/v＞＜/prd＞	九江【是】江西的北大门 技术出口也【有】了进展 中美【签订】合作协议
＜adv＞	状语	LB	＜adv＞【LB】【XP-SBJ】＜/adv＞	【被外部世界】广泛关注
＜cmp＞	补语	DER	＜cmp＞【右兄弟节点】＜/cmp＞	各项工作做得【更好】
＜pp＞	介词位	P BA LB	＜pp＞＜p＞【P】＜/p＞＜/pp＞ ＜pp＞＜p＞【BA】＜/p＞＜/pp＞ ＜pp＞＜p＞【LB】＜/p＞＜/pp＞	【在】讲话中 【把】注意力转向其他市场 【被】警方抓个正着
＜ff＞	方位词位	LC	＜ff＞＜f＞【LC】＜/f＞＜/ff＞	大地震【后】
＜un＞	助词位	ETC	＜un＞＜u＞【ETC】＜/u＞＜/un＞	各种税率【等】优惠
＜uu＞	助词位	DEC DEG DER DEV	＜uu＞＜u＞【DEC】＜/u＞＜/uu＞ ＜uu＞＜u＞【DEG】＜/u＞＜/uu＞ ＜uu＞＜u＞【DER】＜/u＞＜/uu＞ ＜uu＞＜u＞【DEV】＜/u＞＜uu＞	重要【的】意义 两国【的】关系也十分友好 里头学问大【得】很 专心【地】工作
＜uv＞	助词位	AS SB MSP SP	＜uv＞＜u＞【AS】＜/u＞＜/uv＞ ＜uv＞＜u＞【SB】＜/u＞＜/uv＞ ＜uv＞＜u＞【MSP】＜/u＞＜/uv＞ ＜uv＞＜u＞【SP】＜/u＞＜/uv＞	我参加【了】救援 【被】淘汰 两岸【所】做出的贡献 我们的心是连接在一起【的】
＜cc＞	连词位	CC CS	＜cc＞＜c＞【CC】＜/c＞＜/cc＞ ＜cc＞＜c＞【CS】＜/c＞＜/cc＞	店铺【和】民房 【如果】我们那样做

2.2　句式的转换规则

句式结构树库所采用的“句式”术语为吕叔湘先生的定义，即“句子的结构格式和结构类型”“特定句式中成分、词类或特征词序列具有相对固定的结构层次和位置顺序”[7]。短语结构树表明短语与短语之间如何组成句子，对于整个句子属于哪种句式没有说明，这是树库标注体系所反映的语法体系与树库构建者的标注理念的区别所导致的。但每种句式都“有章可循”，根据各类句式的标志性特征，如特定的词类、词序列可以判断是否为相应的句式。具体转换规则见表 3。

表 3 句式转换规则

句式标记	含义		转换规则
APP	同位		【XP-APP】<cc fun="APP"/>【兄弟节点】
COO	并列		【NP_1】<cc fun="COO"/><c>连词</c></cc>【NP_2】
SYN	合成谓语	系动词+VP	<prd><v>【VC】</v></prd><cc fun="SYN"/>【VP】
		助动词+VP	*
SER	连动		*
UNI	联合谓语		【VP_1】<cc fun="UNI"/><c>连词</c></cc>【VP_2】
PVT	兼语		*

注：标 * 号的规则见 2.2 节具体说明。

合成谓语句：合成谓语句是 1956 年《暂拟汉语教学语法系统》(以下简称《暂拟系统》)提出的一类句式，句式结构树库做了一些修改，将合成谓语分成“助动词+VP”或“系动词+VP”两类，在句式结构树库中统一标为“<cc fun="SYN"/>”。对于“系动词+VP”这一类，可以利用 CTB 的词性标签“VC”(系动词)，助动词则都标为“VV”，利用词性信息的方法不可行。因此我们根据固定的短语树的层级结构转换，具体为：

① (VP(VV)(VP(VV)))

如果符合该短语树结构，那么两个谓核(VV)间为合成谓语关系。

兼语句：CTB 对兼语结构的描写同样有固定的结构层次，因此总结出“节点标记+树结构层次”的特定层级框架即可提取出句式信息。CTB 将动词分成三类：VE(有)、VV(普通动词)、VC(系动词)，兼语句的提取需制定两条规则：

① (VP(VE)(XP-OBJ(XP-SBJ)(VP)))

如果短语树符合这一结构，则 XP-SBJ 后增加兼语结构标记<cc fun="PVT"/>；

② (VP(VV)(NP-OBJ)(IP)))

如果短语树符合这一结构，则 NP-OBJ 后增加兼语结构标记<cc fun="PVT"/>；

连动句：在 CTB 中，连动结构以如下树结构体现：

① (VP(VP)(VP))

当 VP 下有两个 VP 子节点时，在 VP 的左子树【LP】和右子树【RP】之间增加属性标记<cc fun="SER"/>。

2.3 特殊情况的转换

从短语结构树库转为句式结构树库涉及两者标注体系的不同，但“它们主要描述的都是句法结构，在更深层次上具有一致性”[14]。因此，对于只涉及两者标注体系差异的，根据标记的对应关系即可直接转换，对于涉及标注体系所遵循的语法理论差异的，需要进行特殊处理。例如，CTB 中区分了两种“被”，词性标签分别为 LB 和 SB，LB 用于“NP_0+LB+NP_1+VP”结构中，引出施事；SB 用于“NP_0+SB+VP”结构中，直接接动词短语。在句本位语法中，SB 处理成助词，LB 与其引出的施事作为句子的状语成分。对于前者，本文根据词性的对应关系可以简单处理，后者作为复杂情况进行相应的处理。另外，CTB 非叶子节点上的标签组合关系分两种情况，第一种是常见的“短语标记+功能标记”，例如，“NP-SBJ”，另一种是两个及以上的标签进行组合，例如“IP-IMP-TTL-PRD”，第二类情况也需当作特殊情况处理。

2.3.1 限制性转换方法

在实际转换过程中，CTB 中的某些标签并不能完全对应句本位语法中的句法成分，需要增加限制条件。例如，“PP(介词短语)在汉语中既可作状语，也可以充当补语成分，当介词短语在动词短语之前时为状语，在动词后面时为补语。此类情况较多，部分举例见表 4。

2.3.2 多标签组合的转换方法

多标签组合指三个及以上的标签以“-”连接为一个标签。上文中列举的转换规则所依据的标签，均为典型的“短语标记”、“短语标记-功能标记”或“词性”的组合形式。然而 CTB 的标签种类丰富，每个标签集的标签数量众多，尤其注重对短语功能的描写，因此还存在部分以“短语标记-多个功能标记”为组合形式的标签。

表 4　限制性转换规则

CTB 标记	条件	句法成分	句式结构标记
QP-EXT	左有 VV/VC/VE 兄弟节点	宾语	＜obj＞
	左无 VV/VC/VE 兄弟节点	补语	＜cmp＞
PP	右有 VP 兄弟节点	状语	＜adv＞
	左有 VV 兄弟节点	补语	＜cmp＞

总体来说，对应到句式结构树库的有效标记一般居于组合尾部，尽管多标签组合类型多样，但出现频率低，因此针对多标签组合的转换流程可以简单处理为一个条件循环，而不会产生过多的不可预期的问题，具体如下：判断组合中最后一个标记是否符合核心转换方法或辅助转换方法，如果不符合，则判断前一个标记，直到符合转换条件为止。以“NP-OBJ-SBJ-PN”为例，该标签中“PN”不符合两种转换方法，因此对前一个标记“SBJ”进行判断，其符合核心转换方法中的“主语”转换规则，最后将“NP-OBJ-SBJ-PN”下的子树转换成“主语”。

2.4　词性转换方法

从词性标记的数量来看，句式结构树库的词性标记有 15 个，CTB 词性标记有 33 个，类别更多。我们通过句式结构树库和 CTB 的词性的映射关系直接转换。这种转换方法使得词性的粒度变粗了，但并不会丢失词性的大类信息。词性对应关系详见表 5。

表 5　词性转换规则

句式结构树库		宾州中文树库		句式结构树库		宾州中文树库	
标记	词性	标记	词性	标记	词性	标记	词性
n	名词	NN	普名	d	副词	AD	副词
		NR	专名	p	介词	P	介词
		FW	外来词			BA	“把”,“将”
		URL	网页链接			LB	“被”,“给”
		NN-SHORT	略缩普名				
		NR-SHORT	略缩专名	c	连词	CC	并列连词
t	时间词	NT	时名			CS	从属连词
		NT-SHORT	略缩时名	u	助词	DEC	标句词“的”
r	代词	PN	代词			DEG	所有格“的”
		DT	限定词			AS	体标记
f	方位词	LC	方位词			DER	得
m	数词	CD	基数词			DEV	地
		OD	序数词			ETC	等
q	量词	M	量词			MSP	“所”、“以”等
						SB	“被”
						SP	句尾小品词
v	动词	VV	其他动词	e	叹词	IJ	感叹词
		VC	系动词				
		VE	“有”	o	拟声词	ON	拟声词
a	形容词	VA	表语形容词	w	标点	PU	标点
		JJ	区别词/紧缩形容词				

2.5 树库转换算法

根据前文所述的转换规则，我们设计了短语结构树库向句式结构树库转换的算法，详细流程见算法 1。

算法 1 宾州中文树库向句式结构树库的转换算法

输入：短语结构树 psTree

输出：句式结构树 ssTree

```
nodeSeq←preOrder(psTree)          //先序遍历
for i=0→n do
  node ← nodeSeq[i]
  if node∈Clause then             //若为复句则执行小句切分规则
    clauseRules(nodeSeq[i])
  end if
  if node∈Syntax then             //判断是否符合句法成分转换规则
    syntaxRules(node)
  else if node∈Struct then        //判断是否符合句式转换规则
    structRules(node)
  else
    specRules(node)               //执行特殊转换规则
  end if
  posRules(node)                  //执行词性转换规则
  ssTree←node
end for
```

算法首先对输入的短语结构字符串进行多叉树的数据结构构建，若句子中存在小句(Clause)，首先利用小句的转换规则(clauseRules)进行小句切分。按照先序遍历的方式，对每个节点根据转换规则进行转换，包括句法成分的转换方法(syntaxRules)、句式的转换方法(structRules)和特殊情况的转换方法(specRules)，以及词性的转换方法(posRules)。

3 实验

我们用前文所述的规则完成了 CTB5 共 18 244 句的句式结构转换，并人工标注了 CTB 部分句子作为测试集对转换结果进行评估。为了进一步验证树库转换方法的有效性，我们另外设置了两组实验：基于句式结构自动分析器生成树库的实验、短语自动句法分析器结合转换规则生成树库的实验。

3.1 数据集与评价指标

为了使评测结果更具参考性，我们采用 Liu 和 Zhang[19] 对 CTB5 的数据切分方式，将文件编号 271 至 300 部分作为测试集，并去除了测试集中的 30 句新闻电头，因为这类句子在 CTB 中的结构扁平，无可利用的标记，超出了规则处理的范围，最终人工标注共 318 句。我们对测试集进行了句式结构标注①，以评估三种方法的性能。在标注过程中，分词仍然按照源树库 CTB 的标注规范，词性和句法层面依照目标树库句式结构树库的规范。

3.2 基于人工标注的自动转换结果评估

我们通过句式结构树库标签的分类来看三种方法的具体表现，分别是句子、成分、虚词位与句式标签，具体数据见表 6。

通过数据可以发现：

(1) 树库整体转换结果较好，转换规则对于小句切分、句子主干成分、附加成分的处理不错。小句切分准确是后续句法分析的前提，句子成分是句子分析的最重要的一步，数据说明最后构建的新树库质量较好。

(2) 虚词位的转换 F_1 值最高。虚词为封闭集，在转换规则中，两个树库的词性标签为直接映射关系，转换难度较低。附 NP 的助词位根据 CTB 的词性标记“ETC”转换，但标句词短语 CP 中的“的”有些是结构助词，有些是附 NP 的助词位，由于助词位对实际句法分析与句子理解的影响较小，并考虑到规则复杂性，我们的处理办法是根据对应数量的占比映射到句式结构树库的词类体系，这里将其处理为结构助词，这也导致了附 NP 的助词位召回率较低。句式结构树库与 CTB 的连词集合并不能完全对应，部分连词在 CTB 中标为副词，如作为关联词的“但”，词性标注的不一致导致连词的召回率较低。

(3) 兼语结构、连动结构在句式的转换结果中效果最好。两者是根据特定的 CTB 的短语树结构进行转换的，说明树结构能基本对应相应的句式结构，经过规则的转换结果错误率低。

① 人工标注数据 http://github.com/blcuicall/ctb2stb/data

表 6　基于人工标注的自动转换结果评估　　（单位：%）

大类	小类	P	R	F_1	大类	小类	P	R	F_1
句子	小句	93.81	93.37	93.59	虚词位	助词位（定状补）	93.43	100.00	96.60
成分	主语	91.67	92.17	91.92		助词位（附 NP）	100.00	66.67	80.00
	谓语	92.30	91.33	91.81		助词位（附 VP）	94.07	87.40	90.61
	宾语	91.75	91.40	91.57	句式	并列	96.89	63.39	76.64
	定语	74.84	72.18	73.49		同位	69.23	79.41	73.97
	状语	93.11	93.23	93.17		合成谓语	95.29	77.14	85.26
	补语	85.71	85.71	85.71		联合谓语	72.46	72.46	72.46
虚词位	介词位	95.87	99.69	97.74		兼语	97.22	92.11	94.59
	连词位	93.10	75.00	83.08		连动	90.00	90.00	90.00
	方位词位	99.08	99.08	99.08					
整体精确率							90.68		
整体召回率							88.79		
整体 F_1 值							89.72		

（4）合成谓语结构精确率高，但召回率稍低。经过观察，合成谓语结构的主要问题在于“是字句”，CTB 对于该结构有两种表现形式：((VC)(VP))、((VC)(NP-PRD))，前一种形式可以直接对应合成谓语结构，但后一种形式大部分为动宾结构，少部分为合成谓语结构，规则根据多数情况处理为动宾结构，这可能是导致召回率稍低的原因。

（5）并列和联合成分有时并不以连词连接，顿号以及逗号也能起到连接的作用，如“华侨、华人艺术家”“一个有利可图，有钱可赚的投资环境”，在我们制定的转换规则中，主要利用的是非叶子节点的信息，处理这类并列结构和联合谓语结构需要利用到叶子节点信息，因此规则未完全覆盖此类情况。另外，联合谓语结构“有的通过关联词语（连词或关系副词）突显，有的则依赖 VP 自身的语义逻辑”[7]，表 3 所示的联合谓语结构转换规则暂时只考虑了以连词连接的情况，对于以关系副词和 VP 间的语义逻辑突显的联合谓语结构，如“深刻却又舒缓”“污染严重治理无望”等一般处理为连动结构，这可能是联合谓语结构转换效果不佳的另一原因。

（6）CTB 与句式结构树库对同位短语的定义不同，前者定义更宽泛，如“同等优先、适当放宽的原则”，CTB 也分析为同位结构，因此召回了一些本该是定语的错误样本，定语转换问题也在于此。

为了进一步考察本文提出的树库转换规则的效果，我们进行了两组实验以作比较。

3.3　对比实验

3.3.1　实验设置

实验一：通过训练自动句法分析器自动生成树库是扩建树库的一种通用的方式，但是这种方式对自动句法分析器性能要求较高，具有一定挑战性。目前句式结构自动句法分析器处在初步研究阶段，本文借鉴 Kitaev 等人[20]提出基于自注意力机制的神经网络模型训练了句式结构自动句法分析器。

模型采用编码器-解码器架构，将预训练模型 BERT 用于编码器阶段，将词性、位置作为辅助信息传入模型，编码器对词表征$[w_1,\cdots,w_n]$、词性表征$[m_1,\cdots,m_n]$及位置表征$[p_1,\cdots,p_n]$加和获得词嵌入，随后使用多头注意力机制对词嵌入进行编码，解码器采用 CKY[21-23]算法获得句式结构句法树。数据集来源于北京师范大学构建的句式结构树库，借鉴 Liu 等人[19]的切分方式，对数据集采用 50∶1 方式构建训练集和开发集，如表 7 所示。该实验主要探索句式结构自动句法分析器在树库构建方面的效果。

实验二：首先通过短语结构自动分析器产生短语结构，然后利用短语结构树库向句式结构树库的

自动转换算法产生句式结构，这种方式不仅可以扩充更大规模的句式结构树库，而且相较于实验一，不依赖于句式结构树库作为训练数据，具有更强的适用性。

短语结构自动句法分析技术目前较为成熟，常用的短语句法分析器有伯克利句法分析器①、CoreNLP②，但是这些模型输出的短语结构仅有短语标签，如“NP”“VP”等，并无功能标签，如“SBJ”“OBJ”等。如前文所述，本文提出的转换规则需要用到 CTB 的多种标签，特别是功能标签，因此上述句法分析器对于本文提出的自动转换算法并不适用。鉴于此，本文借鉴 Kitaev 等人[20,24]提出的基于自注意力机制的神经网络方法，训练得到可以分析功能标签的短语句法分析器。数据集采用 Liu 等人[19]的切分方式，如表 7 所示。

表 7 句式结构树库和宾州中文树库数据统计

数据集	训练集	开发集	测试集
句式结构树库	67 558	1 352	—
宾州中文树库(CTB5)	17 544	352	318

3.3.2 实验结果与分析

表 8 列出了基于句式结构自动句法分析的方法、基于短语结构自动句法分析结合转换规则的方法和基于宾州中文树库结合转换规则的方法的总体性能和分别在小句、成分、虚词位和句式等方面的性能。

表 8 三种方法对比实验结果 （单位：%）

方法	F_1 值				精确率	召回率	F_1 值
	小句	成分	虚词位	句式			
句式结构自动句法分析	95.21	80.61	88.74	**87.44**	83.85	85.01	84.43
短语结构自动句法分析＋转换规则	95.09	86.01	94.33	74.03	88.40	86.73	87.56
宾州中文树库＋转换规则	**95.70**	**88.43**	**95.72**	78.87	**90.68**	**88.79**	**89.72**

从表 8 可得以下结论：

本文提出的基于宾州中文树库结合转换规则的方法整体效果最优，相比基于句式结构自动句法分析的方法③ F_1 值高出 5.29%，说明基于规则的转换算法在树库自动构建上具有一定优势。在具体标签类别上，转换规则在小句切分、句子成分、虚词位上的效果均优于另外两种方法，在句式上低于句式结构句法分析器，通过 3.2 节的分析，原因在于规则对同位、并列、联合谓语结构的处理存在不足，这是需要继续完善的部分。

短语结构自动句法分析结合转换规则的方法，相较于句式结构自动句法分析的方法，精确率高 4.55%，F_1 值高 3.13%。在成分和虚词位上的 F_1 值分别高 5.40%、5.59%，进一步说明转换规则的有效性。

综上所述，本文提出的基于宾州中文树库结合转换规则的转换方法在句式结构树库构建上具有一定优势，基于短语结构自动句法分析结合转换规则的方法由于不依赖现有人工标注的宾州中文树库，因此在构建大规模的句式结构树库上具有更强的通用性。

4 结语

本文以宾州中文树库为源树库，通过基于规则的方法实现了向句式结构树库的自动转换，以此构建了大规模的新闻领域句式结构树库，并基于人工标注的评估，验证了该方法的有效性。

此外，本文设置了对比实验，以比较句式结构自动句法分析、短语结构自动句法分析结合转换规则、基于转换规则这三种方法的性能。实验表明，基于转换规则的转换方法优于其他两种方法，进一步验证了转换规则的有效性。

目前我们的转换规则仍然存在一些不足，如合成谓语、并列、联合等转换规则还有待完善。

新树库仍然保留源树库 CTB 的分词，未来我们将继续完善句法层面的转换规则，并探索词法层面的转换，以提高新句式结构树库的质量，为自动句法分析等相关研究提供有效的数据支持。

参考文献

[1] XUE N, XIA F, CHIOU F D, et al. The Penn Chinese Treebank: Phrase structure annotation of a large corpus[J].

① https://parser.kitaev.io/

② https://corenlp.run/

③ 此方法性能较低的部分原因是宾州中文树库和现有的句式结构树库分词标准不一致。

Natural Language Engineering，2005，11(2)：207-238.
[2] 周强. 汉语句法树库标注体系[J]. 中文信息学报，2004，18(4)：2-9.
[3] LUCIEN T. Éléments de syntaxe structural[M]. Paris：Klincksieck，1959.
[4] HE W，WANG H，GUO Y，et al. Dependency based Chinese sentence realization[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP，2009：809-816.
[5] 彭炜明，宋继华，王宁. 基于句式结构的汉语图解析句法设计[J]. 计算机工程与应用，2014，6：11-18.
[6] 朱德熙. 语法讲义[M].北京：商务印书馆，1982.
[7] 彭炜明，句本位语法的中文信息处理理论与实践[M]. 北京：外语教学与研究出版社，2021.
[8] LIN D K. A dependency-based method for evaluating broad-coverage parsers[C]//Proceedings of the 14th International Joint Conference on Artificial Intelligence，1995：1420-1425.
[9] XIA F，PALMER M. Converting dependency structures to phrase structures[R]. Pennsylvania Univ Philadelphia，2001.
[10] ŽABOKRTSKÝ Z，SMRZ O. Arabic syntactic trees：From constituency to dependency[C]//Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics，2003：183-186.
[11] NIU Z Y，WANG H，WU H. Exploiting heterogeneous treebanks for parsing[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP，2009：46-54.
[12] KONG L，RUSH A M，SMITH N A. Transforming dependencies into phrase structures[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2015：788-798.
[13] 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)：22-28.
[14] 李正华，车万翔，刘挺. 短语结构树库向依存结构树库转化研究[J]. 中文信息学报，2008，22(6)：14-19.
[15] 周惠巍，黄德根，钱志强，等. 短语结构到依存结构树库转换研究[J]. 大连理工大学学报，2010 (4)：609-613.
[16] 张引兵，宋继华，彭炜明，等. 短语结构树库向句式结构树库的自动转换研究[J]. 中文信息学报，2018，32(5)：31-41.
[17] 石定栩. 乔姆斯基的形式句法：历史进程与最新理论[M].北京：北京语言文化大学出版社，2002.
[18] 黎锦熙. 新著国语文法[M].长沙：湖南教育出版社，2007.
[19] LIU J，ZHANG Y. Shift-reduce constituent parsing with neural lookahead features[J]. Transactions of the Association for Computational Linguistics，2017，5：45-58.
[20] KITAEV N，CAO S，DAN K. Multilingual constituency parsing with self-attention and pre-training[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics，2019：3499-3505.
[21] COCKE J. Programming languages and their compilers：Preliminary notes[M]. New York University，1969.
[22] YOUNGER D H. Recognition and parsing of context-free languages in time n3[J]. Information and Control，1967，10(2)：189-208.
[23] KASAMI T. An efficient recognition and syntax-analysis algorithm for context-free languages[J]. Communi Cations of the ACM. 1970，13(2)：94-102.
[24] KITAEV N，KLEIN D. Constituency parsing with a self-attentive encoder[C]//Proceedings of the 56th Annual meeting of ACL，2018：2676-2688.

谢晨晖(1998—)，硕士研究生，主要研究领域为自然语言处理和智能语言学习。
E-mail：xch15673171321@163.com

杨麟儿(1983—)，通信作者，博士，副教授，主要研究领域为人工智能、自然语言处理、计算机辅助语言学习。
E-mail：lineryang@gmail.com

胡正升(1995—)，硕士研究生，主要研究领域为自然语言处理和智能语言学习。
E-mail：hzs0828@163.com