北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2016.005

基于规则的依存树库错误自动检测与分析

史林林　邱立坤†　亢世勇

鲁东大学文学院，烟台　264025；† 通信作者，E-mail: qiulikun@pku.edu.cn

摘要　尝试将依存树转化为短语结构树，并基于规则的方法自动检测出人工标注结果中的错误。将该方法应用于已经过两遍人工校对的北京大学多视图依存树库，从 50275 个句法树中发现 1529 处错误，且正确率为 100%。进一步，所有错误可以分为 3 个层次：分词错误、词性与句法角色不符、句法角色错标。该方法可以有效提高依存树库的质量，并且适用于各类型的依存树库。

关键词　树库；词性；句法角色；错误检测

中图分类号　TP391

Rule-Based Detection and Analysis of Annotation Errors in Dependency Treebank

SHI Linlin, QIU Likun†, KANG Shiyong

School of Chinese Language and Literature, Ludong University, Yantai 264025;
† Corresponding author, E-mail: qiulikun@pku.edu.cn

Abstract　The authors try to transform dependency tree into phrase structure tree, and detect annotation errors automatically based on manual rules. The method has been used in processing Peking University Multi-view Chinese Treebank (PMT). Although PMT has been manually checked twice before processed by this method, 1529 errors were detected among the 50275 sentences and the precision is 100%. The errors mainly belong to three types: word segmentation error, mismatching between POS and syntactic role, and syntactic role error. This method can further improve treebank quality, and be applied to other dependency treebanks.

Key words　treebank; part of speech; syntactic role; error detection

树库是在分词和词性标注的基础上对句子中词与词之间句法关系进行标注所形成的语料库。近年来，树库作为训练和评价统计句法分析器的数据基础，越来越受到研究者们的重视[1]。在被用来训练统计句法分析器时，树库质量对句法分析器效果有较大影响，因此提高树库质量是一个重要的研究课题。目前已有较多学者探索使用统计方法自动检测人工标注的树库中存在的错误[2−5]。此类方法用于辅助人工进行第二遍校对，在一定程度上可以降低工作量，提高工作效率。经过两遍人工校对后的树库仍然可能存在一定的错误，其中有许多是因为词性和句法两个层面标注不协调造成的，也可能是因为标注人员偶然误操作导致的。本文把经过两遍人工校对后的依存树库作为处理对象，试图找出人工标注结果中的错误，以进一步提高树库质量。

本文提出一种基于产生式规则的错误检测方法，其基本原理是从依存树转换到短语结构树时，如果生成短语功能范畴失败，则通常是因为依存树标注错误所导致。Rambow[6]认为，短语结构树和依存树只是两种不同的句法表现形式，在表达能力上并没有高下之分：一般地，短语结构树中标注有短语功能范畴标记和层次信息，依存树中标注有中心语和语法角色信息；但事实上在短语结构树中也可以标注中心语和语法角色信息，在依存树中也可

国家自然科学基金(61572245, 61103089, 61272215)资助
收稿日期: 2015−06−19；修回日期: 2015−08−15；网络出版时间: 2015-09-30 12:35:34

以标注短语功能范畴标记和层次信息。如果在其中任何一种形式的句法树中同时标注了中心语、语法角色、层次和短语功能范畴标记信息，则一定可以无歧义地向另一种句法树转换。本文主要考察从常规依存树生成常规短语结构树中的短语功能范畴的过程，并在这一过程中基于产生式规则自动检测出人工标注错误，进而对错误进行分析，给出各类型错误的分布。本文方法理论上适用于各种类型的依存树库，但是在具体实施时需要为所处理的依存树库构建一套产生式规则，这套规则涉及词性、依存关系类型和短语功能范畴。

1 基于短语功能推导的错误检测

1.1 短语功能的可推导性

对于短语整体功能的可推导性，汉语学界很早就有过讨论。朱德熙[7]提出: “内部构造相同的结构，功能一般相同；功能相同的结构，内部构造不一定相同。” 陈保亚[8]则将之总结为结构功能原则，“如果两个言语片断的直接成分功能相同，结构关系相同，它们的功能也相同。”“这个规律叫做结构功能原则。根据这一原则，只要知道了直接成分的功能和结构关系，结构功能就知道了”。换言之，如果已知具有依存关系的两个词的词类以及它们之间语法关系的类型，就可以推导出这两个词所构成的短语的整体功能。

按照结构功能原则，推导短语整体功能时在每一步都需要知道直接成分的功能。依存树中只标注有词的功能标记，没有短语的功能标记，但是通过递归的方式，可以依次获得各短语直接成分的功能标记。

本文使用短语功能标记作为推导的目标标记，推导短语整体功能的规则为: 父结点词类+子结点词类+语法角色=>短语整体功能标记。比如“v+n+VOB=>VP”表示父结点词类为动词(v)、子结点词类为名词(n)、子结点充当父结点的宾语(VOB)，则整个短语的功能类型为动词性短语 VP。

1.2 依存句法体系

本文采用北京大学多视图依存树库(Peking University Multi-view Chinese Treebank, PMT)[9]的词性体系和依存句法体系。该词性体系对北京大学 2003 版词性标记集[10]进行简化，包含 33 个词性标记，如名词(n)、动词(v)、形容词(a)、副词(d)、状态词(z)、介词(p)、连词(c)、助词(u)、数词(m)、处所词(s)、人名(nr)、标点符号(w)等；依存句法体系中定义了 30 种句法角色，如表 1 所示。

1.3 推导规则集的建立

每个推导规则可以分为条件和结论两部分，例如“v+n+VOB=>VP”这一规则的条件是“v+n+VOB”，结论是“VP”。因此推导规则集的建立分为如下两个部分。

1) 条件库的自动抽取。给定一个人工检查过的树库，可以很容易地将所有可能的条件抽取出来。具体步骤为: 遍历树库中每一棵树，对树中的每一条弧，获取父结点词性、子结点词性和子结点的依存标签，将三者连接成一个字符串，即可生成一个条件，将条件存储到条件库中即可。

2) 人工填写结论。条件库中的条件可能存在

表 1 PMT 依存句法标记集

Table 1 Dependency category set of Peking University Multi-view Treebank

编号	依存关系类型	符号	编号	依存关系类型	符号	编号	依存关系类型	符号
1	核心	HED	11	时体	MT	21	地字	DI
2	主语	SBV	12	数量补语	QUC	22	得字	DEI
3	话题	TPC	13	定语	ATT	23	重叠	RED
4	强调	FOC	14	数字	NUM	24	独立结构	IS
5	宾语	VOB	15	并列式独立结构	ISC	25	小句	IC
6	间接宾语	IOB	16	数量	QUN	26	标点	PUN
7	行为宾语	ACT	17	前附加	LAD	27	一般并列	COO
8	连动	VV	18	后附加	RAD	28	共享并列	COS
9	补语	CMP	19	介宾	POB	29	同位	APP
10	状语	ADV	20	的字	DE	30	跨句标点	PUS

表 2　推导规则示例

Table 2　Example derivation rules

父结点	子结点	语法角色	功能	父结点	子结点	语法角色	功能	父结点	子结点	语法角色	功能
v	d	ADV	VP	q	m	RAD	QP	s	w	PUN	NP
v	p	ADV	VP	p	p	COO	PP	m	q	QUN	NP
v	v	ADV	VP	p	c	LAD	PP	v	v	IC	IP
l	v	COO	VP	p	n	POB	PP	n	w	PUS	IP
v	v	COS	VP	m	n	APP	NP	r	n	SBV	IP
v	u	MT	VP	n	n	APP	NP	r	v	SBV	IP
v	w	PUN	VP	n	n	ATT	NP	v	nr	VOB	vp
v	w	PUS	VP	n	u	ATT	NP	t	n	SBV	IP
r	r	RED	VP	n	ns	ATT	NP	v	n	SBV	IP
v	n	VOB	VP	n	n	COO	NP	a	n	TPC	IP
v	v	VOB	VP	n	m	IS	NP	a	ns	TPC	IP
q	q	COO	QP	n	n	IS	NP	v	n	TPC	IP
q	v	IS	QP	n	n	ISC	NP	u	n	DE	DNP
q	c	LAD	QP	n	w	PUN	NP	u	v	DE	CP
q	m	NUM	QP	d	p	ADV	ADVP	a	a	COS	ADJP
q	q	QUN	QP	a	w	PUS	ADJP	a	a	RED	ADJP

错误，因此需要进行人工审核。在审核时，将所有条件按照频次降序排列，然后一一判断。如果认为条件可靠，则为之添加一个结论(即短语功能标签)；否则，将之剔除。

表 2 中列出一些高频的推导规则作为示例。

1.4　错误自动检测

使用上面所建立的推导规则，可以自动地生成短语结构语法树所需要的短语功能；如果所遇到的条件是推导规则中所没有的，系统将会认为是一个人工标注错误。具体流程如下。

遍历每一棵依存树:
　遍历每一个结点:
　　假定子结点词性为 P_c，父结点词性为 P_h，子结点依存关系标签为 R，通过字符串连接可得到条件“P_h+ P_c+ R”
　　检索规则库中的条件部分:
　　如果检索成功，则继续处理
　　如果检索失败，则简化条件，将父结点词性和子结点依存关系标签连接成条件“P_h+ R”，重新检索:
　　如果检索成功，则继续处理
　　如果检索失败，则将父结点记为人工标注错误

如上所述，在检测过程中，我们在严格条件匹配失败的情况下放松了条件，使得系统对于训练数据中没有出现过的条件也能够匹配上，具有一定的适应能力；同时也确保了错误检测的高正确率。需要说明的是，如果仅使用严格条件匹配，则可以检测出更多的人工标注错误，但相应的正确率会降低。

2　实验及分析

2.1　实验结果

本文实验数据为北京大学多视图依存树库中的新闻树库[9]。树库文本来自人民日报 1998 年 1 月份前 10 天语料(共计 14000 余句)和 2000 年 1 月份全部语料(总计 50000 多个句子)。在建立规则库时，使用 1998 年 1 月份树库，测试时使用 2000 年 1 月份树库。所有树库均经过两遍校对。

在建立规则库时，自动抽取的条件数为 2279，

经过人工检测后是 843 条规则。被剔除的条件中，有一些是因为频次较低且可以被其他规则所覆盖，有一些属于错误标注。

基于该规则库，使用上述检测方法从测试数据中检测出 1529 处错误，正确率为 100% (自动检测出的错误经人工判定均为真正的错误)。进一步分析发现，标注错误可以分为词语切分、词性标注和句法标注三个层次，每个层次又有若干个小类。各类错误的分布如表 3 所示。

2.2 分词错误

汉语书面表达方式以汉字作为最小单位，词与词之间没有空格或其他分隔标记，因此词语切分成为汉语文本处理中首先要解决的问题。自动词语切分中主要的难题是分词歧义消解和未登录词识别。本文在检测树库标注错误过程中发现，有一些句法标注错误是由于词语切分不当所引起的。此类型错误共有 57 处，占总数的 3.70%，具体又分为组合型歧义和人名两类。

2.2.1 组合型歧义

词语切分歧义一般分为两种：交集型歧义和组合型歧义。对于交集型歧义，可根据字段内部提供的信息或以句法为主的局部上下文信息解决。而组合型歧义，切与不切，则导致分词不同，词性不同，语义不同，如图 1 所示。

“就是”合在一起，有助词、副词、连词三个词性；分开后，则为两个词“就/d 是/v”，是状中结构。图 1 中，“就是”显然为两个词，这样整个句子才会有一个谓语中心，有一个根节点。从依存树向短语结构树转换时，由于规则库中不存在“d+v+VOB”(父结点为“就是”，其词性为 d；子结点为“坚持”，其词性为 v；子结点句法关系标注为 VOB，即宾语)这一条件，检索失败；放松条件后检索“d+VOB”，仍然失败。没有能够生成相应短语的功能范畴，系统中直接显示出父结点的词性“d”(即副词)，进而将之判断为一个标注错误。树库中类似词语包括“就是”、“还是”、“才能”、“只有”等，在人工校对中，应根据语境信息判断该合还是该分。

2.2.2 姓名处理不当

在 PMT 标注体系中，姓与名应合成一个词。实际语料中有少数姓名标注不当，造成错误，如例 1 中的“廉颇”，作为人名，应合在一起(为方便起见，以下例句中用“P”标识目标词的父节点)。

例 1 盛泽田/nr “/w [廉/a_ATT] [颇/d_SBV] 未/d [老/a_P] ”/w

2.3 词性与句法角色标记不符

非兼类词在切分的同时一般就可以确定其词性，兼类词的词性则需要依据上下文语境予以判断。因此词性标注导致的句法标注错误主要由兼类

表 3 错误类型及所占比例

Table 3 Distribution of error types

大类	比例/%	小类	比例/%	类型	比例/%
分词错误	3.70	组合型歧义	2.28		
		姓名处理不当	1.42		
词性与句法角色不符	60.50	时间词与句法角色不符	6.41		
		动词与句法角色不符	6.86	动词错标为介词	3.79
				动词错标为副词	3.07
		形容词与句法角色不符	8.24		
		介词与句法角色不符	4.38	介词错标为连词	3.01
				介词错标为副词	1.37
		成语、简称、习用语处理不当	34.60		
词性正确，句法角色错标	35.80	动宾结构错标为介宾结构	19.95		
		数词修饰动词错标为数字	7.13		
		连词句法角色错标为状语	3.14		
		数量补语(QUC)错标	2.35		
		状中结构与述宾结构混淆	3.20		

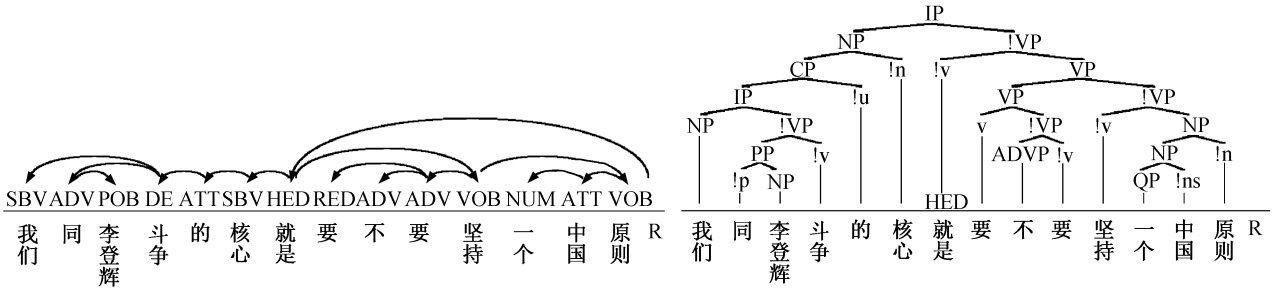

图 1　组合歧义型错误示例

Fig. 1　An instance of Combinatory error

词词性标注不当所导致。此类型错误有 925 处，占总数的 60.50%，包含 5 种类型：时间词与句法角色不符，动词与句法角色不符，形容词与句法角色不符，介词与句法角色不符，成语、简称、习用语处理不当。

2.3.1　时间词与句法角色不符

时间词可作状语，与副词有类似的语法功能。“同期”是指“同一时期”，除作状语外，还可以作宾语、定语，所以作为时间词更合适。“同期”错标为副词，当它充当介词宾语时就会形成“p+ d+ POB”这样的不合法条件，导致向短语结构树的转换失败，如例 2。

例 2　[比/p_ADV] [上年/t_ATT] [同期/d_POB] [增加/v_P] 11.1%/m

2.3.2　动词与句法角色不符

1) 动词错标为介词。

在现代汉语中，单音介词几乎都是从动词发展而来的，单音介词和动词之间有来源关系和大量的兼类现象[11]。语言是动态的，在变化与发展过程中，部分动词虚化为介词。在虚化过程中，有的词虚化比较彻底，完全失去了动词的意义和功能，成为典型介词；而有的还没有失去动词性，有时表现为动词功能，有时表现为介词功能，这部分词属于动介兼类词。因此在标注动介兼类词时，往往会难以抉择，如例 3。

例 3　俄新社/nt 把/p 这/r 篇/q 谈话/n 作为/v 独家/n 新闻/n [提供/v_P] [给/p_CMP] 了/u 《/w 匈牙利/ns 民族/n 报/n》/w。/w

动词的 4 个特征：①能用“不”或“没”否定；②能带宾语；③能带动态助词“着、了、过”；④能重叠。介词保存了动词的前两个特征，失去后两个特征[12]。因此，判断动词和介词时，能加动态助词或者重叠的是动词，反之是介词。同时，动词可单独使用，单独做谓语。例 3 中，“给”后有时态助词“了”，所以应标注为动词 v。否则，就会出现介词与助词“了”搭配的条件，导致转换失败。

2) 动词错标为副词。

此类现象包含两种情况：一是“没有(没)”，动词和副词的区分有误，如例 4；一种是理解判别错误，将动词标为副词，如例 5。

例 4　桃仙/ns 机场/n 已经/d [没有/d_P] 因/p 飞机/n 延误/v 滞留/v 的/u [旅客/n_VOB]。/w

例 5　我/v [决心/d_P] 把/p 我/r 的/u 后半生/n [贡献/v_VOB] 给/p 我国/r 的/u 科学/n 事业/n。/w

“没有(没)”是动副兼类词。否定人物或事情的存在时是动词，否定动作或性状的存在时是副词。例 4 中“没有”用于否定人物“游客”，因此应为动词。例 5 这种情况，属于误标。

2.3.3　形容词与句法角色不符

郭锐[13]提出用“很[不]～〈宾〉”格式把形容词和能受“很”修饰的动词区分开。然而一些词既可以带真宾语(端正态度)，也可以受“很”修饰，不适用于“很[不]～〈宾〉”格式。这些词带真宾语时体现的是“使动”意义，应看作词本身的意义。由于词义不同，受“很”修饰和带真宾语时应看作不同的词，处理为形容词和动词的兼类词。

能否带真宾语则是形容词和动词区分的标准。自动检测中，会将形容词后带宾语的情况当作错误。这种情况下，又分为两种类型：一是动形兼类，将动词标为形容词，却带了宾语，如例 6 中的“落后”；二是形容词词性正确，但直接带了宾语，如例 7。

例 6　在/p 8/m 分钟/q 时间/n 里/f 将/p 比分/n 追/v 成/v 45:46/m，/w 仅/d [落后/a_P] 一/m [分/q_VOB]。

例 7 江苏/ns 玻璃厂/n 的/u 产品/n [走俏/a_P] [市场/n_VOB]

2.3.4 介词与句法角色不符

1) 介词错标为连词。

介词和连词均为虚词: 介词用在词或短语的前面, 构成一个介宾结构, 表示时间、地点、方法、原因等关系; 连词用来连接词语或短语, 表示联合关系或从属关系。“因”兼属介词和连词, 二者意义上有联系, 属于兼类词。在实际标注时, 容易判别错误, 如例 8。

例 8 人们/n [因/c_ADV] 这个/r “/w 新/a 千年/t”/w 而/c[漾/v_P]起/v 无限/z 遐思/n 。/w

“因”在后接名词或名词短语时往往充当介词, 所构成的介宾结构充当状语成分。上述例子中“因”后接名词短语“新千年”, 所以应为介词。

2) 介词错标为副词。

汉语中存在少量介副兼类词, 比如“将”, 需要根据句法功能和语境小心判断, 如例 9。

例 9 [将/d_ADV] 通过/p 资本/n 市场/n 得到/v 的/u [资金/n_POB], /w 集中/a 用于/v 集团/n 战略/n 发展/v 产业/n

“将”为副词时, 表示将要; 作为介词时, 用于引介跟谓词有关的受事。例 9 中, “将”引介跟“用于”有关的“资金”。

2.3.5 成语、简称、习用语处理不当

北大 2003 版词性标记集中有成语、简称、习用语的独立词性标记, 但 PMT 体系中依据语法功能将它们归入相应的词类, 即名归名, 动归动。成语、简称和习用语不是根据句法功能划分出来的词类。由于人民日报语料库中存在一些没有标注小类的成语、简称和习用语, 在进行词性简化时也无法将之归入相应的词类, 因此在进行句法树转换时会导致转换错误, 如例 10, 11 和 12。这类错误也是数量最多的错误类型, 有 529 处, 占 34.6%。

例 10 效果/n 更/d [是/v_P] 如汤沃雪/i 一般/a

例 11 清华/n、/w 北大/n 、/w 对外经贸大/j 、/w 首师大/j 等/u 大多数/m 高校/n 也/d [成立/v_P] 相关/n 领导/n 小组/n

例 12 失业/n 人员/n 只要/c [不挑不拣/l_P] ,/w 保证/v 随时/d 提供/v 就业/n 岗位/n

例 10 成语“如汤沃雪”应归入动词, 例 11 简称“对外经贸大”和“首师大”应归入名词, 例 12 习用语“不挑不拣”则应归入动词。

以上 5 种类型是词性与句法角色不符的错误, 经过分析, 可以得知: 除去误标情况外, 兼类词是最易引起分歧和错误的, 比如动介兼类、动形兼类、介连兼类等, 所以, 在词性标注时应注重兼类词的判别。其次是未处理成语、习用语和简称, 导致出现错误, 此种错误较容易发现和改正。如果准确地分析和判别兼类词, 恰当地处理成语、习用语和简称, 仔细地排除误标情况, 那么依存树库中自动检测出的错误就会减少很多。

2.4 词性正确, 错标句法角色

2.3 节中找到的错误是词性不正确导致的句法角色不符。在自动检测中, 还有一种错误, 即词性正确但句法角色标注错误。此类型错误有 547 处, 占总数的 35.80%。

2.4.1 动宾结构错标为介宾结构

动宾结构和介宾结构是两个区分度较大的结构, 而且语料中已有正确的动词词性, 但标注人员在进行句法标注时忽略了词性, 因此容易将动词宾语 VOB 标记成介宾 POB, 如例 13 和 14。

例 13 [隶属/v_P] 以色列/ns [工党/n_POB] 的/u [罗宾什坦/nr_P]

例 14 未/d [经/v_P] 医师/n [注册/v_POB] 取得/v 执业/n 证书/n

例 13 中的“工党”和 14 中的“注册”, 实际是“隶属”和“经”的宾语 VOB, 但被标为介词宾语 POB。这类现象, 是标注时忽略词性所造成的。

2.4.2 数词修饰动词错标为数字

数词通常跟量词组成数量短语, 然后再做句法成分。但在新闻中领导人讲话时, 会出现“数字+动词”, 这是强调关于动词的几方面内容, 应为状语, 而不是简单地标为数字, 如例 15。

例 15 [四/m_NUM] 到位/v —/w 思想/n 到位/v、/w 感情/n 到位/v、/w 工作/v 到位/v、/w 服务/v 到位/v

2.4.3 连词句法角色错标为状语

根据 PMT 句法标注体系, 连词标为前附加 LAD。上文中提到, 介连兼类时, 标注人员容易混淆二者的语法角色。但此类现象并不是因为词性标注错误而产生的, 反而是因为忽略词性而导致前附加 LAD 错标为 ADV, 如例 16。

例 16 从未/d [因/c_ADV] 接受/v 馈赠/v 而/c 向/p 苏鲁希/nr [提供/v_P] 任何/r 方便 n

2.4.4 数量补语(QUC)错标

数量结构有 4 种语法角色标记: 直接修饰名词, 在名词前面, 做数量短语(QUN); 充当名词的补充

成分，在名词后面，通常定语后置时，做数量补语(QUC)；直接充当谓语动词的右侧子节点标记，为补语(CMP)；充当表示变化(包括增加、减少、改变)词的宾语(VOB)。4 种语法角色易混淆，特别是数量补语和补语的情况，如例 17。

例 17　煤矿/n 企业/n 工资/n 基金/n 平均/a [保持/v_P] 节余/v 6 /m 个/q [月/n_QUC]

数量补语和补语补充说明的对象不同，数量补语针对的是名词，而补语针对谓语动词。因此，例 17 中“6 个月”应作“保持”的补语。

2.4.5　状中结构与述宾结构混淆

状中结构中的修饰语跟中心语会形成种种的语义关系。其中一种表示描写性的，表示动作的变化或情状的变化，可以有两种形式表示：“V(A)+ V”和“V(A)地+ V”。述宾结构前后是支配与被支配、关涉与被关涉的关系。述语主要由及物动词充当，少数由形容词充当，宾语一般是体词或体词性短语，也可以是谓词、谓词性短语。这样两种结构都有“V(A)+ V”形式，导致在判断时会出现错误，如例 18。

例 18　对/p 那些/r [坚决/a_P] [贯彻/v_VOB] 党/n 的/u 路线/n 方针/n 政策/n 的/u 干部/n

2.5　小结

上述错误均为使用本文提出的方法处理已经过两遍人工校对的树库时自动检测出来的，可归为两类。一是句法标注所依据的分词和词性标注结果有误。句法标注工作是基于已有的分词和词性标注结果进行的，因此这部分错误不属于句法标注过程产生的错误。但是通过本文的方法将这些错误检测出来，有助于进一步提升树库的质量。二是在进行句法标注时没有考虑词性和语法角色之间的选择限制关系，凭主观感觉标注语法角色。在进行句法标注时，要将词性和语法角色作为一个整体，既要看词性，又要考虑与之相对应的语法角色，做到词类和句法成分的一致。

3　相关工作

在依存树向短语结构树转化方面，Xia 等[14]对比了 3 种转化算法，仅区分了论元和修饰语，而没有使用依存范畴。Xia 等[15]假设一个既定的依存树与所期望的短语结构树的平面化版本相同，进而提出依存树向短语结构树转化的算法，并且设计了一系列转化规则。他们在错误分析中发现并列结构和标点的错误占转化错误的 32.1%。Bhatt 等(2011) [16]提出 3 种依存树向短语结构树转化的情景分析，Bhatt 等(2012) [17]进一步讨论了转化中的 7 种空语类现象。

4　结语

本文提出一种基于产生式规则的依存树库人工标注错误检测方法，该方法以经过两遍人工校对的依存树库为处理对象，可取得 100%的正确率。由于树库已经过两遍人工校对，存在的错误较少，因此该方法所检测出来的错误数量较少，但这些错误涉及到分词、词性标注结果与句法标注之间的不协调现象，均属于硬伤，修改这些错误对于提高树库质量具有重要意义。该方法适用于各类型依存树库。

参考文献

[1] Abeillé A. Treebanks: building and using parsed corpora. Dordrecht: Kluwer Academic Publishers, 2004, 2: 181

[2] Ambati B, Agarwal R, Gupta M, et al. Error detection for treebank validation // The 9th International workshop on Asian Language Resources (ALR). Chiang Mai, 2011: 23–30

[3] Volokh A, Neumann G. Automatic detection and correction of errors in dependency tree-banks // Proceedings of the 49th ACL: short papers— Volume 2. Stroudsburg, PA, 2011: 346–350

[4] Agarwal R, Ambati B, Sharma D. A hybrid approach to error detection in a treebank and its impact on manual validation time // Linguistic Issues in Language Technology. Stanford University, Palo Alto, CA, 2012, 7(1): 1–12

[5] Agrawal B, Agarwal R, Husain S, et al. An automatic approach to treebank error detection using a dependency parser // Lecture Notes in Computer Science. Berlin: Springer, 2013: 294–303

[6] Rambow O. The simple truth about dependency and phrase structure representations // HLT-NAACL. Los Angeles, 2010: 337–340

[7] 朱德熙. 语法讲义. 北京: 商务印刷馆, 2003

[8] 陈保亚. 20 世纪中国语言学方法论. 山东: 山东教育出版社, 1999

[9] Qiu Likun, Zhang Yue, Jin Peng, et al. Multi-view Chinese treebanking // Proceedings of COLING.

Dublin, 2014: 257–268

[10] 俞士汶, 段慧明, 朱学锋, 等. 北大语料库加工规范: 切分·词类标注·注音. 汉语语言与计算学报, 2003, 13(2): 121–158

[11] 马贝加. 在汉语历时分析中如何区分动词和介词. 中国语文, 2003, 1: 59

[12] 石毓智, 李讷. 汉语语法化的历程. 北京: 北京大学出版社, 2001

[13] 郭锐. 现代汉语词类研究. 北京: 商务印刷馆, 2002

[14] Xia Fei, Palmer M. Converting dependency structures to phrase structures // Proceedings of HLT. Toulouse, 2001: 1–5

[15] Xia Fei, Rambow O, Bhatt R, et al. Towards a multi-representational treebank // LOT Occasional Series, volume 12. Netherlands Graduate School of Linguistics, 2008: 159–170

[16] Bhatt R, Rambow O, Xia Fei. Linguistic phenomena, analyses, and representations: understanding conversion between treebanks // Proceedings of IJCNLP. Chiang Mai, 2011: 1234–1242

[17] Bhatt R, Xia Fei. Challenges in converting between treebanks: a case study from the hutb // Proceedings of META-RESEARCH Workshop on Advanced Treebanking, in conjunction with LREC. Istanbul, 2012: 1–8