多领域中文依存树库构建与影响统计句法分析 *

邱立坤 [1]，史林林 [1]，王厚峰 [2]

（1.鲁东大学文学院，山东省烟台市 264025；2. 北京大学计算语言学研究所，北京市 100871）

摘要：为提升依存分析并分析影响其精度的相关因素，该文构建了大规模中文通用依存树库和中等规模领域依存树库。基于这一系列树库，通过句法分析实验考察质量、规模、领域差异等因素对中文依存分析的影响，实验结果表明：（1）树库规模和质量均与句法分析精度成正相关关系，质量应先于规模因素被优先考虑；（2）通用树库和领域树库之间的差异程度与前者对后者的替代性成相关关系；（3）两种树库混合使用的效果同样与领域差异有关。

关键词：依存树库；领域迁移；依存句法分析

中图分类号：TP391　　**文献标识码**：A

Construction of Multi-Domain Chinese Dependency Treebanks and Analysis of Influencing Factors on Dependency Parsing

QIU Likun[1], SHI Linlin[1] , WANG Houfeng[2]

(1. School of Chinese Language and Literature, Ludong University, Shandong 264025, China;
2. Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract: To boost Chinese dependency parsing and analyze influencing factors on Chinese dependency parsing, we constructed a large-scale general treebank and several middle-scale treebanks for specific domains. Using these treebanks, we performed experiments to evaluate the influence of treebank quality, scale and domain difference upon parsing accuracy. Experimental results show that both treebank quality and scale are positively related to parsing accuracy, and quality is more important than scale. Our experiments also demonstrate that general treebanks and domain treebanks are complementary. In addition, whether a general treebank and domain treebank should be used together is also dependent on the difference between them.

Key words: dependency treebank; domain adaptation; dependency parsing

1 引言

依存句法分析的目标是为给定句子中的每个词找出一个合适的父结点，并标记子结点与父结点之间的句法关系，它是目前最常用的句法分析理论之一。作为主流依存分析方法的统计句法分析，通常用包含大量依存句法树的树库作为训练数据，采用基于图的方法 [1]或基于转移的方法 [2]训练，可得到面向新闻文本的高质量自动句法分析器。依存句法分析已在机器翻译、自动问答、情感分析等领域得到广泛应用，可在一定程度上提升相关系统的性能。但是，统计句法分析性能依赖于树库的规模、质量，并且表现出领域相关性，在迁移到新领域时精度急剧下降 [3]。

* **收稿日期：**　　**定稿日期：**

基金项目：国家社科基金重大项目（12&ZD227）；国家自然科学基金青年项目（61103089，61370117）；山东省优秀中青年科学家科研奖励基金（BS2013DX020）；鲁东大学人文社会科学研究项目（WY2013003）

目前已经有一些文献研究树库转换和融合[4-5]、自学习方法[3]等提高句法分析精度并改善领域迁移效果，但是受语料类型和规模的限制，中文方面很多问题没有得到深入分析。首先是树库规模问题。目前已有一些研究考察树库规模对句法的影响[6]，但使用的树库量级仅在 1 万句左右，本文将考察树库规模增加到 5 万甚至 10 万句时的句法分析效果。其次是树库质量问题，目前尚未见到这方面的研究。最后是通用树库与特定领域树库融合的问题。在中文分词和词性标注上有少量类似研究[7]，句法分析层面暂无。

为考察上述问题，我们基于统一的依存句法标注体系，构建了大规模（12.8 万句）的中文通用新闻树库和中等规模（从 1.7 到 4 万句不等）的特定领域树库。其于这些树库，本文设计了系列实验，以分析树库规模、质量和领域差异对句法分析尤其是特定领域句法分析精度的影响。

本文组织如下：第 2 节介绍依存树库的标注体系、构建流程、所构建树库的基本信息，并简单分析各树库之间的差异；第 3 节通过系列实验分析质量、规模和领域差异等因素对句法分析精度的影响；第 4 节介绍相关工作；最后一节是结论。

2 多领域依存树库的构建

2.1 依存句法标注体系

表 1 PMT 依存体系

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

依存树库的构建必须遵循一定的标注体系，标注体系的差异首先表现在依存关系标签的设置上。各种依存标注体系采用的依存关系标签数量差别较大，标签的内涵更是大不相同。就中文而言，目前有四种体系：（1）由宾州短语结构中文树库转换而来的依存树库（简称 CTB），标签数量为 12 个[1]。（2）哈工大依存体系（简称 HTB），初始版本为 24 个标签，目前版本为 14 个标签[8]。（3）北京大学多视图树库依存体系（简称 PMT）[9]，含 30 个句法标签，该体系参考了 CTB 和 HTB，其中一些标签专门为由依存树转换为短语结构树而设置。

[1] http://w3.msi.vxu.se/~nivre/research/Penn2Malt.html

（4）斯坦福依存体系，该体系标签数量最为庞大[10]。第（1）和（4）体系均依据手工制定的规则生成，不存在直接依据该体系构建的原生树库。

标注体系的差异还表现在对同一句法现象的不同处理策略上。比如，CTB 将兼语句等同于小句宾语句，HTB 和 PMT 则将之以类似于双宾句的方式处理，并设置了专门标签将之与双宾句区别开来。又如，CTB 区分了主语和话题，PMT 也继承了这一做法，用以处理汉语的主谓谓语句；HTB 则允许一个动词带多个主语，不对主语和话题进行区分。再如，CTB 没有显式标注并列结构，因此其依存体系并没有表示并列的标签；HTB 设置了并列标签，且以左结点为核心结点；PMT 设置了并列标签，且以右结点为核心结点。其中，并列结构的处理方式对依存弧方向影响最大，因而也是导致各家树库依存弧差异的主要原因。

本文工作所使用的树库均基于 PMT 体系构建，该体系所使用的依存关系标签如表 1 所示。PMT 体系的特点在于，以依存语法体系为基础，预先考虑了从依存语法到短语结构语法转换过程中的歧义消解问题，因此标注一套依存语法树库（标注依存弧和依存关系标签）可同时得到一套短语结构树库（推导出层次和短语范畴）[9]。

2.2 语料选择与构建流程

表 2 多领域树库基本信息一览

领域	规模（句）	平均句长（词）	平均依存距离	未登录词比例		
				V1	V2	V3
通用	128738	24.0	3.82	8.7	4.5	3.5
医药	32227	19.7	3.51	30.8	21.5	19.4
口语	40620	8.9	2.34	12.0	6.4	5.3
专利	17035	32.2	4.37	30.0	19.5	17.7
微博	29840	13.7	2.93	16.1	9.0	7.7

本文构建的树库包括新闻、医药、口语、专利、微博五个领域，各领域句子数和平均句长如表 2 所示。新闻语料含有政治、科技、社会、教育、体育等多个子领域和叙述文、散文、报告文学、说明文等多种文体，可称之为通用树库；相应地，可称其它树库为领域树库。

新闻树库的文本来自 1998 年 1 月份 1 到 10 日共 10 天语料、2000 年 1 月全部语料、2000 年 2 月全部语料、2000 年 3 月前 20000 句语料，总计 128738 句。其中，1998 年 1 月（14463 句）和 2000 年 1 月（50275 句）经过两遍校对，剩余语料仅经过一遍校对。为表述方便，我们将 1998 年 1 月树库称为 V1（12000 句，不含用于开发和测试的 2463 句），V1 加上 2000 年 1 月树库后称为 V2（62275 句），V2 加上 2000 年 2 月和 2000 年 3 月前 20000 句树库后称为 V3（126275 句）。

医药领域语料来自皮肤病领域教材和论文摘要，口语领域语料来自对外汉语口语教材，专利领域语料来自中文专利文献，微博领域语料为随机抽选的微博，这四个领域树库仅经过一遍校对。

进行一遍校对时，参与人员通常在 10 到 20 人之间。进行二遍校对时，参与人员比一校人员经验更为丰富，人数通常在 4 到 6 人之间。所有树库均按照 PMT 体系的标注规范、采用相同的流程、使用相同的辅助工具构建。

2.3 多领域树库差异分析

不同领域的树库在词汇和语法等层面存在明显差异，我们可以用平均句长、未登录词比

例、平均依存距离等指标来度量领域差异。句长指的是每个句子所含词语的数量。依存距离指的是依存树中子结点与父结点之间所间隔的词的数量，其最小值即子结点与父结点相邻时的值为 1[11]。未登录词指的是出现在测试文本中但未出现在参照文本中的词语，未登录词比例指的是测试文本中未登录词数量占其总词数的比例；显然，当参照文本不同时，未登录词比例也会有所不同。

表 2 中列出了通用树库和四个领域树库的规模等信息，计算平均句长和平均依存距离时以整个树库为计算范围；计算未登录词比例时分别选择 2463 句、1000 句、1000 句、1000 句、1000 句、1000 句为各领域的测试文本（分别来自 1998 年 1 月人民日报树库的最后位置和四个领域树库的最后位置），分别选择 V1、V2、V3 三个版本的通用树库作为参照文本，从而计算出三种未登录词比例。

如表 2 所示，平均句长与平均依存距离具有明显的相关性，句长值越大，依存距离也越大。CTB 上的实验 [12]表明同一领域的句子，句长值越大，则句法分析的精度越低。但是句法分析受到多种因素的影响，不同领域之间的句长与句法分析精度之间并没有必然联系。

从 V1、V2 到 V3，随着参照文本规模的增大，各树库未登录词比例相应减少。比较之下，口语和微博两个领域未登录词比例要远远低于医药和专利两个领域。如果以未登录词比例为衡量领域差异的标准，则可以认为口语和微博两个领域与通用新闻领域差异较小，医药和专利两个领域与通用新闻领域差异较大。

3 影响统计句法分析精度因素之分析

基于所构建的大规模通用树库和中等规模的领域树库，可以分析质量、规模和领域差异等因素对句法分析精度的影响。

3.1 实验设置

数据 对于通用树库，参照 Qiu 等 [9]选择 1998 年 1 月份树库的 12001-13000 句作为开发集合，13001-14463 句作为测试集合(由于二校版本质量更高，因此在所有相关实验中，通用新闻树库均选择二校版本作为测试数据)。对于四个领域树库，各选择最后的 1000 句作为测试集合。

依存句法分析器 本文在训练和测试时使用 MATE-tools 依存句法分析器 3.61 版 [2][13]。该句法分析器支持多线程训练，在多核计算机上可以获得较高的训练速度；在精度上与 ZPar[14] 等句法分析器相当 [9]，处于领先水平，明显优于 MaltParser 和 MSTParser[15]。

评测标准 在评价依存句法分析精度时，我们使用 UAS（Unlabeled Accuracy Score）和 LAS（Labeled Accuracy Score）两个指标。UAS 指不考虑依存关系标签时依存弧标注正确的结点数占总结点数的比例，LAS 指同时考虑依存关系标签和依存弧时标注正确的结点数占总结点数的比例。后续实验中在没有特别说明的情况下均使用 UAS 值进行比较，LAS 值仅做参考。

3.2 树库质量

在人工校对树库时，二校人员从一校人员中选拔而来，其熟练程度、对规范的把握程度均明显优于一校人员；二校在一校基础上进行，其主要工作为修改一校人员校对结果中的错误。因此一般情况下二校结果在质量上优于一校结果。表 3 列出了 V1、V2 和 V3 三个树库的一校、二校版本用做训练数据时的句法分析精度。

[2] https://code.google.com/p/mate-tools/

在同等规模的情况下，二校树库均明显优于一校树库。在使用 V1、V2 和 V3 时，二校比一校分别提升 0.87%、1.36%和 1.02%。值得特别说明的是，二校 V2 规模仅为一校 V3 的一半，精度却高出 0.47%。这一结果充分说明树库质量对句法分析精度有较大影响，对一批树库进行两遍校对所得到的句法分析器精度上可能优于对两倍规模的树库进行单遍校对所得到的句法分析器。较小的树库规模意味着占用内存较小和运行速度更快，因此在规模和质量间平衡时，应优先考虑质量。

表 3 基于一校、二校树库的句法分析精度比较

质量	规模	UAS(%)	LAS(%)
一校	V1	84.82	81.56
	V2	87.13	83.89
	V3	88.02	84.79
二校	V1	85.69	82.82
	V2	88.49	85.68
	V3	89.04	86.06

3.3 树库规模

句法分析精度与用做训练数据的树库规模关系也非常密切。表 3 反映了三种不同规模的新闻树库句法分析精度的差异，V2 规模是 V1 的 5 倍，V3 规模是 V2 的 2 倍。从表 3 可以看出，无论一校树库还是二校树库，在树库规模增大时，句法分析精度均有明显上升，从 V1 到 V2，两种版本的 UAS 分别提升了 2.3%和 2.8%，此时树库规模扩大了 4 倍；从 V2 到 V3，UAS 分别提升了 0.89%和 0.55%，此时树库规模扩大了 1 倍。二校版本中从 V2 到 V3 的提升低于一校版本，主要原因是二校版本中 V3 相比于 V2 增加的树库并没有经过二校。下文在没有特别说明时，V1、V2、V3 均指其二校版本。

表 4 基于不同规模特定领域树库的句法分析精度比较

	1000		2000		5000		10000		全部	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
医药	81.17	77.64	82.60	79.35	84.94	81.80	85.19	82.19	87.03	84.13
口语	80.77	76.55	82.86	78.97	85.30	81.80	87.25	83.78	89.49	86.41
专利	75.31	72.21	77.50	74.50	79.14	76.16	84.90	82.62	84.99	82.66
微博	74.61	68.76	76.64	71.10	78.55	73.37	79.79	74.57	82.09	76.95

表 4 反映了不同规模的特定领域树库句法分析精度上的差异，规模从 1000、2000、5000、10000 到全部树库。从该表可以看出，在所有领域中，当树库规模增加时，句法分析精度逐渐提高。比较之下，医药、口语和微博三个领域规模与精度增加的趋势较为一致；专利领域树库从 5000 增加到 10000 时，句法分析精度提升幅度明显比其它三个领域大，规模进一步增加时句法分析精度基本上没有新的提升。导致这一差异的主要原因是专利文献包含化工、电子、机械、医药等多个子领域，子领域之间差异较大，从 5000 增加到 10000 时所增加的语料与测试语料比较接近，因此带来较大幅度的提升。具体而言，在 5000 句时，医药、口语、专利、微博四个领域测试数据的未登录词比例分别为 8.1%、12.1%、9.9%和 8.4%；增大到 10000 句时，未登录词比例分别降为 6.7%、8.5%、3.7%、6.3%。其中专利领域未登录词比例降幅最大，这应该是导致专利领域精度显著上升的主要原因。这一结果说明，对于专

利这样的复杂领域，应考虑对子领域进行细分，对各子领域分别建立语料库。

3.4 领域差异

表 5 基于通用树库的句法分析器在四个领域上的句法分析结果

	V1		V2		V3	
	UAS	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**
医药	78.23	74.16	80.28	76.80	81.27	77.90
口语	85.34	79.14	88.93	82.79	88.81	82.32
专利	73.13	69.52	74.54	70.93	74.61	71.21
微博	76.09	70.43	79.54	73.96	79.64	73.79

为考察领域差异对句法分析的影响，我们进行了两种实验：其一是测试基于通用树库训练的句法分析器在特定领域树库上的句法分析精度；其二是测试基于通用树库加一定数量领域树库训练的句法分析器在领域树库上的句法分析精度。前一种实验的结果如表 5 所示，用作训练数据的通用树库包括 V1、V2 和 V3 三个版本，相应地在每个领域树库上可以得到三个句法分析结果。从该表可以看出，从 V1 到 V2 各领域的句法分析精度均有稳定提升，幅度从 1.4%到 3.6%；从 V2 到 V3 时，医药领域有 1%左右的提升，但口语、专利、微博三个领域仅有微小提升甚至有所下降。

基于通用树库的最优句法分析效果在医药（81.27%）和专利（74.61%）这两个领域中基本与使用 1000 句领域树库训练的结果（分别为 81.17%和 75.31）相当（参见表 4）；在口语和微博这两个领域中则可与使用 10000 句领域树库训练的结果相当。如表 2 所示，医药和专利这两个领域与通用新闻的差异较大，未登录词比例在 17%以上；口语和微博这两个领域则与通用新闻差异较小，未登录词比例在 8%以下。由此说明，在与通用新闻差异较小的领域中，通用树库对领域树库的替代性[3]较好，当领域树库规模较小时，其性能通常会弱于通用句法分析器，因此没有必要构建小规模的此类树库；在与通用新闻差异较大的领域中，通用树库对领域树库的替代性较差，有必要为特定领域构建新的树库。

如 2.3 节所述，领域差异体现在多个角度（平均句长、平均依存距离、未登录词比例等），上述实验表明以未登录词比例为标准的领域差异与领域迁移时句法分析精度变化的趋势呈现明显的相关性，因此在后续的分析中主要使用未登录词比例作为度量领域差异的标准，未登录词比例越高，则领域差异越大。

表 6 基于通用树库 V1 加领域树库的句法分析器在四个领域上的句法分析结果

	V1+1000		V1+2000		V1+5000		V1+10000		V1+全部	
	UAS	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**	**UAS**	**LAS**
医药	81.90	78.55	82.86	79.81	**84.26**	**81.30**	85.58	82.54	**87.10**	**84.22**
口语	87.08	81.27	87.62	82.10	87.90	83.16	88.98	85.08	90.53	87.22
专利	76.59	73.75	78.21	75.42	**79.05**	**76.33**	85.17	82.99	**84.89**	**82.64**
微博	78.31	72.94	79.20	73.75	79.75	74.61	80.65	75.36	82.24	77.17

后一种实验的结果如表 6 和表 7 所示。表 6 中通用树库为 V1（12000 句），领域树库的规模包括 1000、2000、5000、10000 和全部五种。当领域树库规模为 1000 时，通用树库

[3]如果使用前者训练的句法分析器精度上好于基于后者训练的句法分析器，或者与后者相当，则我们认为前者对后者的替代性较好，否则可认为替代性较差。

加领域树库的效果明显好于单独使用通用树库（参见表 5）或者领域树库（参见表 4）的结果，说明此时通用树库和领域树库的互补性较强。当领域树库规模为 5000、10000 和全部时，这一趋势基本未变，但医药和专利两个领域中效果有所减弱，通用树库加领域树库的效果基本与单独使用领域树库相当甚至比之稍差。这一结果说明，当领域树库达到一定规模（比如 5000 句以上）且与通用领域树库差异较大时，可单独使用领域树库训练句法分析器，其精度与领域树库加上通用树库相当；当与通用领域树库差异较小时，混合使用通用和领域树库训练的句法分析器通常能比单独使用领域树库有一定程度的提升。

表 7 基于通用树库 V2 加领域树库的句法分析器在四个领域上的句法分析结果

	V2+1000		V2+2000		V2+5000		V2+10000		V2+全部	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
医药	**83.07**	**79.70**	83.53	80.43	83.98	81.02	84.85	81.81	86.45	83.66
口语	**89.24**	**83.30**	89.01	83.24	89.19	83.39	89.49	84.26	90.48	86.58
专利	**76.95**	**73.93**	77.92	74.77	78.84	75.88	84.31	81.76	84.52	81.96
微博	**81.02**	**75.53**	80.92	75.34	81.24	75.83	82.09	76.81	82.42	77.24

表 7 中通用树库为 V2（62275 句），领域树库的规模包括 1000、2000、5000、10000 和全部五种。当领域树库规模为 1000 时，通用树库加领域树库的效果明显好于单独使用通用树库（参见表 5）或者领域树库（参见表 4）的结果，并且好于表 6 中的相应精度；规模为 2000 时，口语和微博两个领域比规模为 1000 时有所下降，医药和专利两个领域则继续上升；规模为 5000、10000 和全部时，精度均继续上升，但是医药和专利两个领域均比表 6 中的相应精度要低。该结果表明，当领域树库规模较小（2000 以下）时，通用树库规模越大，与领域树库混合使用时所取得的提升也越明显；当领域树库规模较大（5000 以上）时，通用树库规模的持续增大，并不一定能带来精度提升，当通用树库和领域树库领域差异较大时甚至会带来少量下降。

4 相关研究

中文树库方面，目前达到一定规模的中文树库有宾州短语结构树库（CTB）[16]、Sinica 依存树库 [17]、清华短语结构树库 [18]、国家语委短语结构树库 [19]、北大短语结构树树库 [20] 和哈工大中文依存树库（HTB）[8]，其规模分别为 160 万词（2013 版 [21]）、36 万词、100 万词、100 万字、130 万词、111 万词。就文本类型来说，CTB 包括新华社新闻、新闻杂志、博客、广播访谈、广播新闻等多种类型，HTB 主要来自 1992 年到 1996 年人民日报，清华树库分新闻、文学、说明文、科技四种语体。

树库转换和融合方面，李正华等 [4]将 CTB 转换成 HTB，并混合起来进行句法分析实验，在加入小规模 CTB 时，句法分析精度有所提升，进一步增加时则有所下降。Li 等 [5]提出新的转换方法，将 HTB 转换为 CTB，并混合起来进行实验，在 CTB5 和 CTB6 上分别提升了 1.37%和 1.10%。两个研究的结论有所不同，可能的原因是后者采用了新的转换方法提升了转换质量。从 CTB5 到 CTB6 提升的幅度有所下降，主要是因为 CTB6 的规模（78 万词）大于 CTB5（51 万词），从而使得新加入树库（HTB）的影响变小。

此外，Sagae 等 [6]分析了树库规模对句法分析的影响，实验中使用的树库（英文树库 GENIA，内容为生物学科技文献摘要）规模从 100、200 一直到 1000（以 100 为间隔），之后从 2000、3000 一直到 8000（以 1000 为间隔），实验结果表明在 1000 句之间，每增加 100 句都会有显著提升，1000 之后每增加 1000 句也只会有缓慢提升。这一结果与本文规模因素

部分（3.3 节）的实验基本一致。与之相比，本文这一方面的实验涉及领域更多、树库规模更大，同时观察到少量异常情况，并用领域差异对之进行了解释。

5 结语

本文基于所构建的大规模通用依存树库和中等规模的领域依存树库，通过一系列实验分析了树库质量、规模和领域差异等因素对中文句法分析精度的影响。实验结果表明：（1）树库质量对句法分析精度有较大影响，对一定规模树库进行两遍校对所得句法分析器性能优于对两倍规模树库进行单遍校对，因此在质量和规模间进行平衡时应优先考虑质量；（2）无论是通用树库还是领域树库，在规模增加（从 1000 句到 12 万句）时均能带来精度的提升，但提升幅度逐渐减少；（3）在已有大规模通用树库的情况下，如果一个特定领域与通用领域差异较小，则没有必要为之构建中等规模（5000 以下）的树库；当特定领域与通用领域差异较大时，即使构建 1000 句规模的树库，性能也可能超过单独使用通用树库；（4）特定领域树库规模较小（2000 句以下）时，混合使用通用树库和领域树库通常能带来明显的提升，此时通用树库规模的增大也能带来进一步的提升；（5）特定领域树库规模较大（5000 句以上）时，如通用树库和领域树库差异较小，则混合使用二者能带来精度提升，如差异较大，则单独使用特定领域树库即可获得与混合使用相当乃至更好的效果。

参考文献

[1] Ryan McDonald, Fernando Pereira, Kiril Ribarov, et al. Non-projective dependency parsing using spanning tree algorithms[C]//Proceedings of HLT-EMNLP, 2005: 523-530.

[2] Joakim Nivre. 2006. Inductive dependency parsing[M]. Springer.

[3] Slav Petrov and Ryan McDonald. Overview of the 2012 Shared Task on Parsing the Web[C]//Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language, 2012.

[4] 李正华,车万翔,刘挺.2008.短语结构树库向依存树库转化研究[J].中文信息学报, 22(6): 14-19.

[5] Zhenhua Li, Ting Liu, Wanxiang Che. Exploiting multiple treebanks for parsing with quasisynchronous grammars[C]//Proceedings of ACL, 2012: 675-684.

[6] Kenji Sagae, Yusuke Miyao, Rune Sætre, et al. Evaluating the Effects of Treebank Size in a Practical Application for Parsing[C]//ACL 2008 Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing, 2008: 14-20.

[7] Meishan Zhang, Yue Zhang, Wanxiang Che, et al. Type-Supervised Domain Adaptation for Joint Segmentation and POS-Tagging[C]//Proceedings of EACL, 2014: 588-597.

[8] Wanxiang Che, Zhenghua Li, and Ting Liu. Chinese Dependency Treebank 1.0 LDC2012T05[DB]. Web Download. Philadelphia: Linguistic Data Consortium, 2012.

[9] Likun Qiu, Yue Zhang, Peng Jin, et al. Multi-view Chinese treebanking[C]//Proceedings of COLING, 2014: 257-268.

[10] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, et al. Discriminative reordering with Chinese grammatical relations features[C]//Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation, 2009: 51-59.

[11] 刘海涛. 2008.基于依存树库的汉语句法计量研究[J]. 长江学术, 2008, 3:120-128.

[12] Wenliang Chen, Jun’ichi Kazama, Kiyotaka Uchimoto, et al. Improving Dependency Parsing with Subtrees from Auto-Parsed Data[C]//Proceedings of EMNLP, 2009, 2: 570-579.

[13] Bernd Bohnet. Top accuracy and fast dependency parsing is not a contradiction[C]//Proceedings of Coling,

2010: 89-97.

[14] Yue Zhang and Stephen Clark. Syntactic Processing Using the Generalized Perceptron and Beam Search[J]. Computational Linguistics, 2011, 37(1): 105-151.

[15] Wanxiang Che, Valentin Spitkovsky, Ting Liu. A comparison of Chinese parsers for Stanford dependencies[C]//Proceedings of EACL, 2012: 11-16.

[16] Nianwen Xue, Fei Xia, Fu-Dong Chiou, et al. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus[J]. Natural Language Engineering, 2005, 11(2): 207-238.

[17] 陈凤仪,蔡碧芳,陈克健,等. 中文句结构树资料库 (Sinica Treebank)的构建[J]. Computational Linguistics and Chinese Language Processing, 1999, 4(2): 87-104.

[18] 周强.2004.汉语句法树库标注体系[J].中文信息学报, 2004, 18(4): 1-8.

[19] 靳光瑾,肖航,富丽,等.现代汉语语料库建设及深加工[J].语言文字应用, 2005, 2: 111-120.

[20] 詹卫东.树库在汉语语法辅助教学中的应用初探[J]. Journal of Technology and Chinese Language Teaching, 2012, 3(2): 16-29.

[21] Nianwen Xue, Xiuhong Zhang, Zixin Jiang, et al. 2013. Chinese Treebank 8.0 LDC2013T21[DB]. Web Download. Philadelphia: Linguistic Data Consortium.

作者简介：邱立坤（1979——），男，博士、副教授，主要研究领域为计算语言学。Email: qiulikun@gmail.com。史林林（1990——），女，硕士研究生，主要研究领域为语料库语言学。Email: shilinalive@163.com。王厚峰（1965——），男，博士、教授，主要研究领域为语篇分析、语言知识库与领域知识库、情感分析等。Email: wanghf@pku.edu.cn。


邱立坤


史林林


王厚峰