文章编号：1003-0077(2019)02-0034-09

面向多领域多来源文本的汉语依存句法树库构建

郭丽娟，彭 雪，李正华，张 民

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：为了支持汉语句法分析研究，目前句法分析领域已经标注了多个汉语依存句法树库。然而，已有树库主要针对较规范文本，而对各种网络文本如博客、微博、微信等考虑较少。为此，该文基于近期研制的标注规范及可视化在线标注系统，开展了大规模数据标注。聘请了 15 名兼职标注者，并采用严格的标注流程保证标注质量，目前，已经标注了约 3 万句的汉语依存句法树库，其中包含约 1 万句淘宝头条文本。该文重点介绍了数据选取、标注流程等问题，并详细分析了标注准确率、一致性和标注数据的分布情况。未来将继续对多领域多来源文本进行标注，扩大树库规模，并以合适的方式公开相应的标注数据。

关键词：依存句法；树库构建；多领域多来源文本

中图分类号：TP391 **文献标识码**：A

Construction of Chinese Dependency Syntax Treebanks for Multi-domain and Multi-source Texts

GUO Lijuan, PENG Xue, LI Zhenghua, ZHANG Min

(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China)

Abstract: The existing Chinese dependency treebanks are mainly annotated for canonical texts, and give little consideration to web texts, such as blogs, WeiBo, and WeChat. This paper presents a large-scale tree-bank annotation, based on the recently designed annotation guideline and online annotating system. Altogether 15 part-time annotators are involved and a strict annotation procedure is applied to guarantee the quality. So far, we have annotated about 30,000 Chinese sentences with their dependency syntax trees, including about 10,000 sentences from Taobao headline texts. This paper describes the details in data selection and annotation workflow. We also analyze the annotation accuracy, inter-annotator consistency, and distribution of annotated data.

Keywords: dependency syntax; treebank construction; multi-domain and multi-source texts

0 引言

树库作为标注了词类、句法和语义等各种语言信息的资源，一方面能为汉语句法学研究提供真实、准确的语言知识；另一方面可以用于训练和测试句法分析器[1]。目前学术界成规模的树库主要包括短语结构树库和依存树库两种类型。和短语结构句法相比较，依存句法具有以下优势：①更适合描述汉语中词间关系错综复杂的句法结构，并且通过依存关系标签可以直接表达词语之间的句法语义关系；②形式简单，便于标注；③句法分析结果的存储空间较小[2]。

现有汉语句法树库的文本主要来源于《人民日报》、语文课本、政府白皮书等规范文本。然而在日益发展的互联网中产生了大量的用户生成数据，例如，产品评论、聊天记录、问题答案等网络用语，极大地丰富了汉语文本。使用在现有的传统树库上训练得到的句法分析器处理网络文本时，其分析准确率急剧下降，说明对于数据驱动的分析模型而言，人工标注数据的规模、质量很大程度上影响着分析结果的准确率。这类分析模型表现出明显的领域相关性，在移植到新领域时，性能会急剧下降[3]。

为了解决这类问题，一方面，学术界有很多学者

收稿日期：2018-09-29 **定稿日期**：2018-10-29

基金项目：国家自然科学基金(61876116, 61673289)；江苏省高校自然科学研究重大项目(16KJA520001)

通过研究树库转换[4-5]、树库融合[6-7]等方法来提高句法分析性能，并改善领域移植效果，然而，受到树库类型和规模的限制，汉语方面还未有比较深入的研究[8]；另一方面，便是构建大规模树库以解决此问题。目前在英文网络文本上的树库构建工作已经逐步展开。2012 年谷歌组织面向邮件、博客、问题答案、新闻组、评论五个来源的英文网络文本，标注了小规模评测数据，命名为 Google English Web Treebank[3]。汉语方面，邱立坤等[8]构建了包括新闻、医药、口语、专利、微博五个领域的汉语依存树库。但与英文相比，面向汉语网络文本的依存句法树库构建进展仍相对缓慢。

基于以上的讨论，我们为了提高汉语网络文本的依存句法分析性能，亟需对不同类型的网络文本分别标注一定规模的语料，为后续的研究工作提供支持。基于对汉语依存树库构建技术比较深入的研究，我们研制了一个新的数据标注规范作为指导，在基于浏览器的在线标注系统中，对于多领域多来源的文本进行程序化的标注，构建了一个面向多领域多来源文本的汉语依存句法树库。

1　相关研究工作

表 1 罗列了目前公开的较大规模的汉语句法树库。Sinica 汉语树库由中国台湾中央研究院构建，从现代汉语平衡语料库中抽取句子进行开发并标注[9]。宾大汉语树库(CTB)最初由美国宾夕法尼亚大学发起，目前由布兰迪斯大学薛念文教授等维护和更新，标注了新闻、评论、广播、访谈等语料[10]。北大汉语树库(PCT)由北大中文系逐步建设[11]，标注了语文课本、政府白皮书、新闻等语料。清华汉语树库(TCT)由清华大学周强教授等建设[12]，标注了文学、学术、新闻等语料。哈尔滨工业大学汉语依存树库(HIT-CDT)由哈工大社会计算与信息检索研究中心建设[13]，标注了《人民日报》语料。北大汉语多视图依存树库(PKU-CDT)由北大计算语言学研究所构建，该树库是以依存语法为核心的多视图汉语树库标注体系，标注了新闻、医药、专利等语料[8]。

表 1　目前公开的较大规模的汉语句法树库

树库	发表时间	语法类型	规模
Sinica 汉语树库	1999	信息为本的格位语法	36 万词
宾大汉语树库(CTB)	2000—2013	短语结构语法	162 万词
北大汉语树库(PCT)	2003—2011	短语结构语法	90 万词
清华汉语树库(TCT)	2004	短语结构语法	100 万词
哈工大汉语依存树库(HIT-CDT)	2012	依存语法	111 万词
北大汉语多视图依存树库(PKU-CDT)	2015	依存语法	140 万词

2　汉语依存句法树库的构建实践

2.1　标注规范

我们的目标是面向多领域多来源文本，不断积累，构建大规模的依存句法树库。为达到这个目标，在制定依存关系标签时，我们充分借鉴 HIT-CDT、PKU-CDT 及通用依存树库(universal dependencies，UD)等树库构建的结果；针对规范的新闻文本，以及网络文本中的各种语言现象，例如，频繁出现的谐音字、插入语、重复、大量表情符、标点符号缺失、旧词新意等现象，结合语言学理论，在标注实践中总结规律，不断扩展，最终制定了一个面向多领域多来源文本的汉语依存句法数据标注规范(目前规范已有 60 多页)，作为整个工作的基础。规范的标签集合如表 2 所示。

2.2　数据选取

我们了解到，标注规范的制定一定程度上缓解了标注一致性低的问题，但依存关系标签的多样性和句法的模糊性，仍会导致在树库构建过程中不同标注者的一致性较差，给树库构建带来困难。

构建局部标注树库对此提供了一个新的解决思路。局部标注意味着标注者只需要标注句子中部分词语，增强了标注者的注意力，使得标注者可以更加将精力集中在这些词语中。通过这种方式，不同标注者之间更容易得到一致的标注结果，为了能最大程度地节省标注时间和成本，又能尽可能得到更多的对分析器有用的信息，对模型训练更有帮助，我们选取待标注数据的原则及流程如下。

1) **选取句子中置信度较低的部分词语进行标注**　Dozat 和 Manning[14]提出基于图的神经网络双仿射模型，使用神经网络模型计算一个句子 $\boldsymbol{x}$ 中每条依存弧的分数。我们利用这个模型来得到句法树分数，即句法树分数只包括了从核心词到依存词的依存弧的分数，如式(1)所示。

$$\text{Score}(\boldsymbol{x},\boldsymbol{d};\boldsymbol{w}) = \sum^{(h,m);h\rightarrow m\in d} \text{Score}(h\rightarrow m;\boldsymbol{w}) \quad (1)$$

其中，$\boldsymbol{d}$ 表示依存句法树，$\boldsymbol{w}$ 表示模型参数，Score($h\rightarrow m$)通过神经网络模型计算得到。

该模型使用了 CRF-loss，所以每棵句法树的概率如式(2)所示。

$$p(\boldsymbol{d}\mid\boldsymbol{x};\boldsymbol{w}) = \frac{e^{(\text{Score}(\boldsymbol{x},\boldsymbol{d};\boldsymbol{w})}}{\sum_{\boldsymbol{d}'\in y(\boldsymbol{x})} e^{\text{Score}(\boldsymbol{x},\boldsymbol{d}';\boldsymbol{w})}} \quad (2)$$

其中，$y(\boldsymbol{x})$表示句子 $\boldsymbol{x}$ 所有可能的句法树。

因此，每条依存弧的边缘概率，就是所有包含这条依存弧的句法树的概率之和，如式(3)所示。

$$p(h\rightarrow m\mid\boldsymbol{x};\boldsymbol{w}) = \sum_{\boldsymbol{d}\in y(\boldsymbol{x});h\rightarrow m\in\boldsymbol{d}} p(\boldsymbol{d}\mid\boldsymbol{x};\boldsymbol{w}) \quad (3)$$

Li 等[15]研究了句法分析任务中，基于局部标注数据的主动学习方法，取得了令人满意的结果。借鉴 Li 等[15]的工作，我们根据每个词语的最有可能的一个核心词 $h^0=\text{argmax}_h p(h\rightarrow i\mid\boldsymbol{x})$的边缘概率来衡量每个单词 w_i 的置信度，如式(4)所示。

$$\text{confidence}(\boldsymbol{x},i) = p(h_i^0\rightarrow i\mid\boldsymbol{x}) \quad (4)$$

置信度越低说明依存弧越不确定，所以在之后选取出句子时，选取置信度较低的 $\alpha\%$的词语进行标注，并将这 $\alpha\%$的词语的平均置信度作为句子置信度。

表 2 依存关系标签集合

关系标签	说明	例句	标注结果
root	sentence root(根节点)	我 爱 妈妈	(\$→爱,root)
sasubj-obj	same subject and object(同主语同宾语)	图 1(c)	(建立→建全,sasubj-obj)
sasubj	same subject(同主语)	图 1(c)	(建立→改进,sasubj)
dfsubj	different subject(不同主语)	图 1(c)	(建立→提高,dfsubj)
subj	subject(主语)	我 爱 妈妈	(我←爱,subj)
subj-in	subject inside a subject-predicate predicate (主谓谓语中的内部主语)	他 确实 头 疼	(头←疼,subj-in)
obj	object(宾语)	我 爱 妈妈	(爱→妈妈,obj)
pred	predicate(谓语)	命令 他 扫地	(他→扫地,pred)
att	attribute modifier(定语)	国家 主席	(国家←主席,att)
adv	adverbial modifier(状语)	非常 喜欢	(非常←喜欢,adv)
cmp	complement modifier(补语)	洗 干净 手	(洗→干净,cmp)
coo	coordination construction(并列结构)	鲜花 和 掌声	(鲜花→掌声,coo)
pobj	preposition object(介宾)	在 家 看书	(在→家,pobj)
iobj	indirect-object(间宾)	给 他 书	(给→他,iobj)
de	de-construction("的"字结构)	这 是 他 的	(他←的,de)
adjct	adjunct(附加成分)	我 走 了	(走→了,adjct)
app	appellation(称呼)	老师 ,你 好	(老师←好,app)
exp	explanation(进一步解释)	普京(俄罗斯 总统)	(普京→总统,exp)
punc	punctuation(标点)	我 爱 妈妈 。	(爱→。,punc)
frag	fragment(片段)	你 ,我 ,中国	(你→我,frag;我→中国,frag)

假设一个实际选取任务："从一批未标注数据池 U 中抽取 1 000 句由 5～25 个词语构成的句子组成待标注数据池 U_3，U_3 中每个句子选取 50%的词语进行标注"。

我们用上面这个例子来具体阐述选取局部标注词语的流程：

① 句法分析器分析词语的置信度。对于未标注数据池 U 中每一个句子，使用句法分析器进行句法分析测试，通过句法分析器分析出每个句子中各个词语的置信度，这些带有词语置信度的句子组成数据池 U_1。

② 选取符合句子长度的句子。从带有词语置

信度的数据池 U_1 中选取数据池 U_2，U_2 中每个句子由 5～25 个词语(标点不算词语)组成。

③ 选取一定比例的词语进行标注。

A. 先将数据池 U_2 中的每个句子中的词语置信度从低到高地排序，取前 50%(假设句子有 6 个词语，选取前 3＝6×50%个)的词语的置信度的平均值作为整个句子的置信度，且每个句子选出的 50%词语即为该句的待标注词语；

B. 将 U_2 中的句子按照句子置信度从低到高排序，选取前 1 000 个句子构成待标注数据池 U_3。

在选取数据时要遵循高比例优先选取原则。即我们在选取不同要求下的待标注数据时，仅考虑选取句子中待标注的词语比例，比例越高越优先选取，而与句子长度无关。因为句法分析器分析的置信度越低，意味着该词语的标注难度越高，更需要对这类词语进行高比例的选取并标注。

2) **舍弃相似度过高且置信度较高的句子**

我们按照 1)选出一批新的待标注数据后，还要和自身数据池中其他句子，以及已标注过的数据池中所有句子进行相似度计算，确保将新数据中相似度较高且置信度较高的句子舍弃。这样可以避免重复工作，以减少人力、物力，从而使标注者集中对难度较高的句子进行标注，保证标注的数据的高质量，以及多样性。

以计算句子 a 和句子 b 的相似度为例，具体的相似度计算方法如下：

A. 为了防止句子分词出错带来的影响，将所有句子处理成以 char ＋bichar 为单位，例如："我是中国人。"处理成"我 我是 是 是中 中 中国 国 国人 人 人。。"；

B. 将句子 a 中的 char 和 bichar 构成一个集合 AS，句子 b 中的 char 和 bichar 构成一个集合 BS；

C. 相似度计算如式(5)所示。

$$\text{Similarity} = \frac{|AS \cap BS|}{\min(|AS|, |BS|)} \tag{5}$$

相似度 Similarity 的阈值按实际数据情况来定。

假设句子 a 和句子 b 都是待标注句子，且两者相似度超过设定阈值，则舍弃置信度较高的句子。

假设句子 a 是待标注句子，句子 b 是已标注句子，且两者相似度超过设定阈值，则舍弃句子 a。

3) **加入地雷**

为了更好地提高数据质量，在按照 1)和 2)数据选取原则选取一批新数据后，我们会在新的数据批次中将以前标注过的有答案的句子作为地雷混入。我们放入地雷有两大作用：

① 自动评价标注者的标注情况；

② 进一步检查之前的标注结果，以便提高标注质量。

通过以上 3 个步骤顺序选取出待标注数据，放入标注系统中进行人工标注。

2.3　标注流程

从提高数据质量的目标出发，同时又能最大化减少数据标注管理者的工作，实现大规模数据标注。我们在一个基于浏览器的在线标注系统中进行程序化标注。图 1 给出了标注系统的标注界面。标注前，所有待标注的词语都用方框标记；当一个方框中的词语用弧和标签标注出它的核心词后，该词语的方框会消失；标注者必须标注完所有方框中的词语，才能单击"提交"按钮。这种标注界面的设计主要是为了支持局部标注(同样也适用于完整标注)。

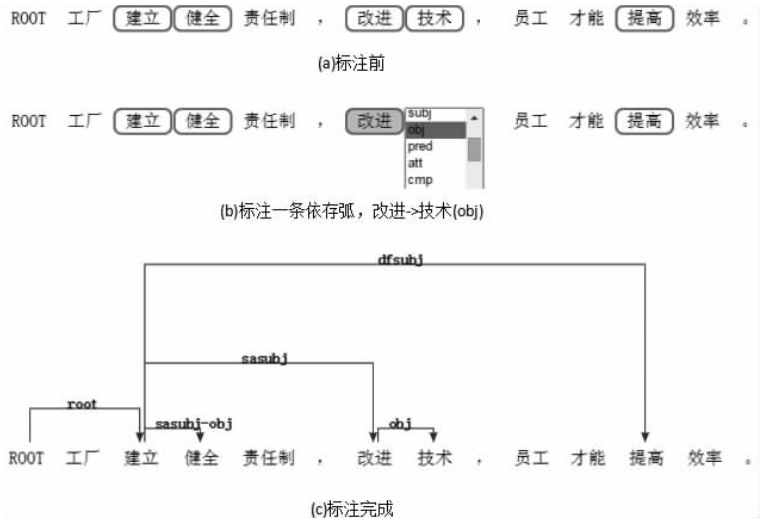

图 1　标注界面

图2给出了一个句子的整个处理流程：

① 标注系统将一个句子随机分配给两个标注者标注。标注完成后，如果两个标注结果完全一致，将答案入库，流程结束。否则进入步骤②。

② 两个标注结果至少有一条弧不一致，就会触发审核机制，系统会将这个句子随机分配给一位专家进行审核，确定唯一答案。进而，标注系统将审核过的答案，反馈给出错的标注者进行学习。学习过程中，如果没有出现投诉，那么就将确定的答案入库，流程结束；否则进入步骤③。

③ 标注人员对答案不认可，提出投诉(若有投诉，我们鼓励标注者多提供投诉理由，以便实现异步沟通，提高数据质量)。系统会将投诉句子随机分配给一位权威专家，确定唯一答案并入库，流程结束。

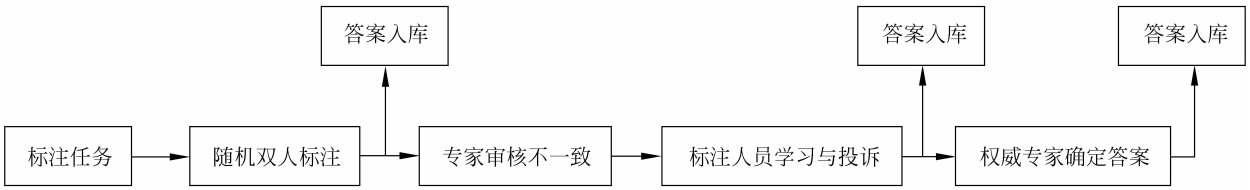

图2 标注流程图

3 标注进展与数据分析

基于以上的树库构建实践工作，我们目前标注了各1万句左右的三个不同领域的依存句法树库，具体数据来源如表3所示。哈工大CDT分为cdt_v1和cdt_v2两批数据批次；淘宝头条 分为content_v1和content_v2两批数据批次；PCTB7数据为一个数据批次 pctb7。对于这些数据，我们按标注时间先后顺序(cdt_v1，content_v1，content_v2，cdt_v2，pctb7)分批次放入系统中标注，分批标注的数据详细信息如表4所示。

表3 数据来源说明表

来源	领域	句子数	总词数(标点词数)	标注词数
哈工大CDT(cdt_v1、cdt_v2)	人民日报，小学课本	10 312	164 562(13 736)	51 022
淘宝头条(content_v1、content_v2)	产品介绍；产品评价；美容、养生、服装搭配等方面介绍	9 043	132 088(17 535)	45 928
PCTB7(pctb7)	新闻(杂志、广播)；对话(广播)；讨论组；博客	11 424	197 866(15 960)	50 209

表4 数据批次信息说明表

数据批次	详细信息(去除病句后)	
	全标注数据	局部标注数据(每句选取不同百分比的词语)
cdt_v1	1 981句，[5,10]词	2 921句，[10,20]词，50%
content_v1	1 933句，[5,10]词	2 564句，[5,20]词，50%
content_v2	—	1 555句，[5,25]词，50%； 2 991句，[5,25]词，20%
cdt_v2	—	5 410句，[5,25]词，20%
pctb7	—	2 447句，[10,25]词，50%； 3 635句，[10,25]词，30%； 5 347句，[10,25]词，20%

标注系统会将一个句子随机分配给两个标注者标注，所以，我们从以下几个方面对每个数据批次进行分析：①所有标注者标注的依存弧平均准确率、一致性及句子一致性；②单个标注者标注依存弧的准确率；③树库标签的分布情况。

其中对于准确率和一致性的计算方法如下：

A. 依存弧的准确率：假设一个句子有5条依存弧需要标注，某个标注者提交的答案中有3条依存弧与最终系统给出的答案相同，则准确率=3/5；

B. 依存弧的一致性：假设一个句子有5条依存弧需要标注，随机分配给a和b两个人标注，两人最终提交的答案(不受他人影响)中一致的依存弧为2条，则一致性=2/5；

C. 句子一致性：假设一个数据批次有10个句子，其中有1个句子被不同的两个标注者标注为完全一致，即需要标注的这个句子中所有依存弧都一致，则句子一致性=1/10。

在计算每个数据批次依存弧的准确率、一致性及句子一致性时，需要注意的两个点是：

① 计算依存弧准确率的分母是计算依存弧一致性的分母的两倍，因为一个句子我们会分配给两个人标注，在计算准确率时，标注者对同一个句子的不同标注结果算作两个不同句子；

② 对于两个人标注一致的依存弧，我们不考虑其正确性，只考虑两人对于一条依存弧的标注理解是否一致。

3.1　整体准确率、一致性分析

对于每个数据批次中所有标注者标注的依存弧，对平均准确率、一致性及句子一致性进行了统计分析，如图 3 所示。结合图 3 和表 4，可以得到以下信息。

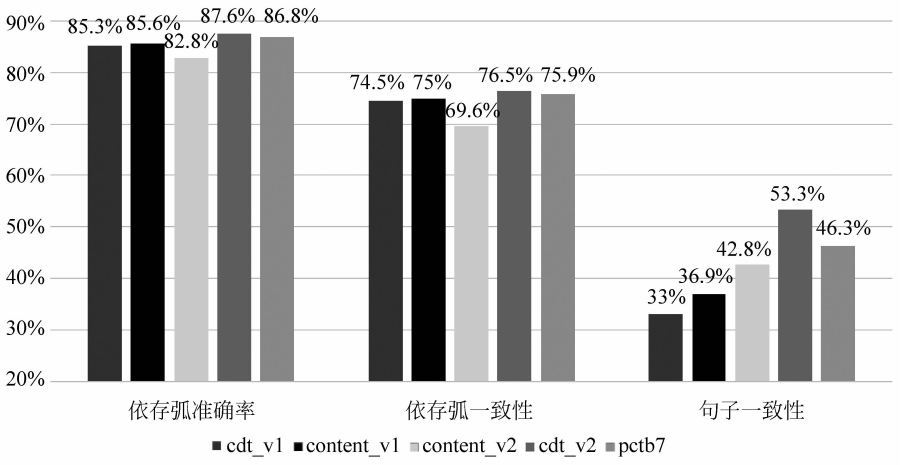

图 3　各批次整体评价指标

(1) 随着标注工作的推进，标注者的标注水平在逐步上升，整体准确率、一致性及句子一致性都有所提高；整体依存弧准确率都在 83%以上，一致性都在 70%以上。说明我们为标注工作制定的规范具有一定的科学性、系统性及完整性，可以充分指导标注者标注。

(2) 单从某个数据批次来看，有以下两点。

① content_v2 中依存弧准确率和一致性都相对较低。原因是：一方面，和前面两批数据相比，content_v2 的句子长度更长，且都是对一定比例置信度较低的词语进行人工标注，一定程度上导致准确率和一致性相对较低；另一方面，和后面两批数据相比，content_v2 的句子长度较短，但其准确率、一致性却依旧较低，充分说明了网络文本数据的标注难度要远大于规范文本数据。另外，content_v2 中句子一致性比前两个数据批次要高，且多是长度较短的句子，说明标注者更容易理解短句子。

② pctb7 中依存弧的准确率、一致性及句子一致性相比 cdt_v2 都有所下降。原因是：一方面，相较于 cdt_v2，pctb7 中的句子长度增加，使得标注难度增大；另一方面，pctb7 中不仅存在需要标注 20%比例词语的句子，而且存在需要标注 50%和 30%比例词语的句子。说明一个句子中需要标注的词语(且这些都是模型分析出置信度较低的词语)增多也会增加标注难度。因此，在之后的数据选取中，尽量不选过长的句子，且一个句子中按比例选取的需要标注的词语个数不得超过某个设定的参数，超出的词语，我们不作为标注任务，以防标注难度过高，从而影响数据质量。

另外，我们对整个树库中标注不一致的标签进行了统计，计算方法为：如果对于某条弧两人标注不一致，则将该弧对应正确的标签数加 1；假设整个数据批次中某个标签正确答案个数为 A，被标为不一致的个数为 B，则不一致性$=B/A$，发现标签不一致性较高的几个为：dfsubj(45.14%)，cmp(37.97%)，sasubj(32.65%)，pobj(30.56%)，coo(29.67%)。

除了以上对新树库数据的弧、标签进行统计分析外，我们也将新标注的树库和原树库进行了比较。但由于新标注树库与原树库是根据不同标注规范进行的标注，标签不易于比较，所以选择对无标签的弧一致性进行统计分析发现：cdt 和原树库 CDT 的无标签弧一致性为 81.58%；pctb7 和原树库 PCTB7 的无标签弧一致性为 66.29%。

3.2　单个标注者准确率分析

由于标注者过多，我们选取的分析对象的标准是：①在 5 个数据批次中至少标注了 3 个数据批次；②标注者在每批数据中标注的依存弧的数目至

少达到1 000条。根据以上标准,我们选取了11位标注者并对他们标准的依存弧准确率进行了统计,如图4所示。

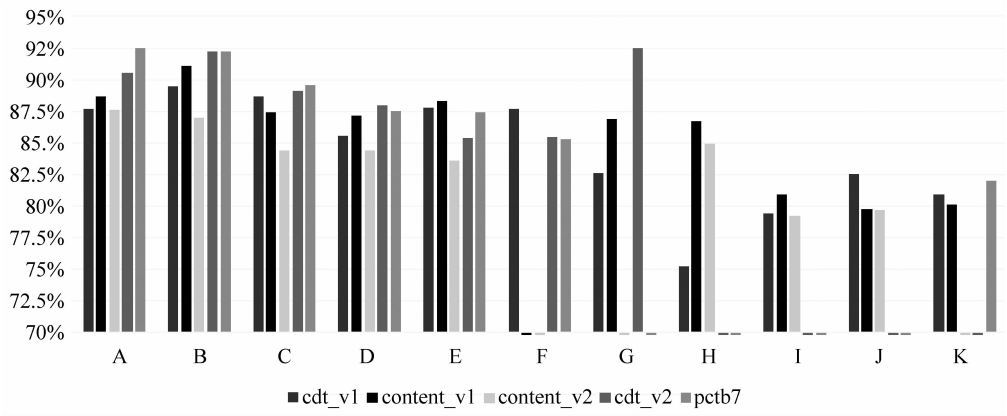

图4 单个标注者准确率

从图4中我们可以得到以下信息:

① 11位标注者随着参与标注工作时间的递增,虽然有所波动,但其依存弧准确率整体呈上升趋势。

② 通过观察发现,前5位参与了所有数据批次标注的标注者的依存弧准确率普遍较高,然而后面只参加了3批数据批次的6位标注者标注的依存弧准确率波动很大并且参差不齐。说明标注工作需要长期坚持,中途一段时间不标注,会对规范有所遗忘,对标注质量有所影响。所以在选取标注者时,应该侧重选择可以长期稳定的标注者,以保证标注质量。

③ 通过分析单个标注者在不同数据批次中标注的依存弧准确率,检验标注者是否能以认真的态度去胜任标注工作。比如标注者J,他的标注数量很少,标注的依存弧准确率低于整体准确率,并且他标注的依存弧准确率随着时间的推移并没有提高,反而有所下降,那么针对这样的标注者,我们会考虑对其重新培训或者辞退,以确保标注数据的质量。

3.3 树库标签分布情况分析

由于较多标签的出现频率很低,所以我们对只有在一个数据批次中数量大于100个的标签进行统计分析。在数据中都分别抽取1 000句全标注数据及局部标注数据(20%),由于pctb7这批数据我们没有进行全标注,所以只抽取了局部标注(20%)数据,标签的分布统计如表5所示。

表5 树库标签分布情况表(%)

标签	1000_all_cdt_v1	1000_all_content_v1	1000_20%_content_v2	1000_20%_cdt_v2	1000_20%_pctb7
subj	15.20	11.00	10.00	9.60	10.20
obj	13.50	12.30	10.30	12.30	12.20
att	23.90	19.70	10.00	27.90	29.10
adv	15.00	21.40	7.60	8.80	13.00
cmp	2.30	1.70	3.00	2.40	1.40
root	14.10	14.30	16.40	8.50	5.70
sasubj	2.00	3.30	22.00	13.40	8.70
dfsubj	0.20	0.50	13.90	5.10	4.90
pobj	3.30	1.40	0.60	3.20	3.30
adjct	9.00	11.80	3.00	3.90	7.70

根据表4和表5分析得到以下信息:

① 无论是在全标注还是局部标注的数据中,用来标注汉语句子中主干(subj(主语)、obj(动宾)、att(定语)、adv(状语)、cmp(补语))关系的标签占比较

大。说明汉语句子中这些主干关系对应的词语置信度整体上都较低，体现了人工标注的重要性。

② 用于标注谓语的 root(根节点)、sasubj(同主语)和 dfsubj(不同主语)这三个标签的占比较大，说明谓词关系在句子中是比较常见的；也能说明从句法角度来看谓词是句子中最重要的词。同时我们可以发现：a)sasubj 和 dfsubj 这两个标签在局部标注数据中的占比远大于在全标注数据中的占比，说明选取局部数据时，这两个标签所对应的词语置信度普遍较低，被大量地选取并标注；b)在 content_v2 的局部标注数据中 sasubj 和 dfsubj 的占比是最大的，说明在长句子的网络文本中谓语句更多。

③ pobj(介宾)和 adjct(附加成分)这两个标签的数量占比较大，原因是：对于 pobj 来说，汉语中的动词和介词理解歧义较大，所以标注者在选择是 obj 还是 pobj 时可能会有一些歧义；对于 adjct 来说，汉语是一种结构化语言，其中有许多只为句子结构完整的助词、叹词等无意义的词语，目前我们都用 adjct 来进行标注，那么在之后的规范更新中都可以将这些着重考虑。另外，pobj 在规范文本中的占比较大，而在网络文本中，助词、叹词等非常常见，所以在 content_v1 的全部标注数据中 adjct 这个标签的占比是最大的。

4　结论与展望

本文介绍了目前我们在面向多领域多来源文本的汉语依存句法树库构建方面所做的一些工作。我们基于前期研制的数据标注规范和在线标注系统，聘请了 15 位标注者，标注了约 3 万句的高质量汉语依存句法数据。本文重点介绍了数据选取、标注流程等问题，对标注数据的质量及标注过程中的一些现象进行了统计分析。

通过这些工作，我们在汉语依存句法树库的人工标注方面积累了一定经验。首先，由于句法标注工作的困难性，我们需要在数据以及标注方面都进行严格的流程控制，以确保标注数据质量；其次，通过统计与分析发现，整体数据的标注弧一致性及句子的一致性都较低，需要审核专家进行进一步的检查，体现了双人标注的重要性。目前我们标注的树库规模还很小，未来我们在现有的树库基础上会进一步构建大规模的面向不同领域不同来源的汉语依存句法树库。

参考文献

[1] 王跃龙，姬东鸿. 汉语树库综述[J]. 当代语言学，2009(1)：47-55.

[2] 李正华. 汉语依存句法分析关键技术研究[D]. 哈尔滨：哈尔滨工业大学博士学位论文，2013.

[3] Petrov S, Google R M, York N, et al. Overview of the 2012 shared task on parsing the web[C]//Proceedings of the 1st Workshop on Syntactic Analysis of Non-canonical Language at NAACL 2012, 2012.

[4] Sato M, Manabe H, Noji H, et al. Adversarial training for Cross-Domain universal dependency parsing [C]//Proceedings of the CONLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2017:71-79.

[5] 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.

[6] Yu J, Elkaref M, Bohnet B. Domain adaptation for dependency parsing via self-training[C]//Proceedings of the 14th International Conference on Parsing Technologies, 2015:1-10.

[7] Li Z, Liu T, Che W. Exploiting multiple treebanks for parsing with quasi-synchronous grammars[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Association for Computational Linguistics, 2012:675-684.

[8] 邱立坤，史林林，王厚峰. 多领域中文依存树库构建与影响统计句法分析因素之分析[J]. 中文信息学报，2015，29(5)：69-75.

[9] Chen K-J, Luo C-C, Chang M-C, et al. Sinica treebank: Design criteria, representational issues and implementation[M]. Abeillè A. [S. l.]: Kluwer Academic Publishers, 2003:231-248.

[10] Xue N, Xia F, Chiou F-D, et al. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus[J]. Natural Language Engineering, 2005, 11(2): 207-238.

[11] 詹卫东. The application of treebank to assist Chinese grammar instruction: A preliminary investigation[J]. Journal of Technology and Chinese Language Teaching, 2012, 3(2): 16-29.

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

[13] Che W, Li Z, Liu T. Chinese dependency treebank 1.0 (LDC2012T05)[DB/OL]. Philadelphia: Linguistic Data Consortium, 2012 http://catalog.ldc.upenn.edu/LPL2012Tos.

[14] Dozat T, Manning C D. Deep biaffine attention for neural dependency parsing[C]//Proceedings of the 5th International Conference on Learning Representations, 2017.

[15] Zhenghua Li, Min Zhang, Yue Zhang, et al. Active

learning for dependency parsing with partial annotation[C]//Proceedings of the 54th Annunl Meeting of the Association for Computational Linguistics, 2016: 344-354.

郭丽娟(1993－),硕士研究生,主要研究领域为句法分析。
E-mail: 335018562@qq.com

彭雪(1994－),硕士研究生,主要研究领域为句法分析。
E-mail: 654905417@qq.com

李正华(1983－),通信作者,博士,副教授,主要研究领域为词法分析、句法分析、语义分析。
E-mail: zhli13@suda.edu.cn

中国中文信息学会 2019 年活动计划

2019 年活动计划表

序号	活 动 名 称	主 要 内 容	时间	地点
1	第十六届自然语言处理青年学者研讨会(YSSNLP2019)	促进青年学者之间的学术交流	5 月 3－5 日	琼海
2	第四届 IEEE 网络空间数据科学国际会议(IEEE DSC2019)	面向数据科学、大数据、网络空间中的数据密集型应用等数据科学和大数据研究热点	6 月 23－25 日	杭州
3	第十四届中国中文信息学会暑期学校(CIPS Summer School)暨《前沿技术讲习班》(ATT)	中文信息处理相关学科的前沿技术讲座	7 月 11－14 日	北京
4	第二届大数据安全与隐私保护学术会议	会议主题“受控共享为数据应用排忧解难,隐私计算为个人信息保驾护航”	7 月 12－14 日	兰州
5	第十七届中国少数民族语言文字信息处理学术研讨会	构建少数民族语言信息处理学术领域的交流平台	8 月 9－11 日	西宁
6	第十五届全国人机语音通讯学术会议(NCMMSC 2019)	人机语音通讯技术领域的研究与开发	8 月 14－17 日	西宁
7	第八届全国社会媒体处理大会(SMP2019)	面向社会媒体的自然语言处理	8 月 16－18 日	深圳
8	第四届语言与智能技术高峰论坛	语言与智能技术论坛(与计算机学会合办)	8 月 24 日	北京
9	全国知识图谱与语义计算大会(CCKS2019)	中文知识图谱构建与应用	8 月 24－27 日	杭州
10	第二十五届全国信息检索学术会议(CCIR2019)	Web 信息检索;事件抽取;文本分类与聚类	9 月 20－22 日	福州
11	第十五届全国机器翻译研讨会(CCMT2019)	机器翻译模型、技术及系统;多种语言机器翻译系统评测	9 月 27－29 日	南昌
12	第十八届中国计算语言学学术会议(CCL2019)中国中文信息学会学术年会及理事会(CIPS2019)	促进中文信息处理领域的理论创新、技术交流与产学研合作;学会 2019 学术年会	10 月 18－20 日	昆明
13	第五届中国健康信息处理学术会议(CHIP 2019)	促进医疗健康文本、图形、图像和生物序列等信息的处理技术的发展	11 月 22－24 日	广州

学会二维码　　学报二维码