文章编号：1003-0077(2018)10-0028-08

适应多领域多来源文本的汉语依存句法数据标注规范

郭丽娟，李正华，彭 雪，张 民

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：近十年来，依存句法分析由于具有表示形式简单、灵活、分析效率高等特点，得到了学术界广泛关注。为了支持汉语依存句法分析研究，国内同行分别标注了几个汉语依存句法树库。然而，目前还没有一个公开、完整、系统的汉语依存句法数据标注规范，并且已有的树库标注工作对网络文本中的特殊语言现象考虑较少。为此，该文充分参考了已有的数据标注工作，同时结合实际标注中遇到的问题，制定了一个新的适应多领域多来源文本的汉语依存句法数据标注规范。我们制定规范的目标是准确刻画各种语言现象的句法结构，同时保证标注一致性。利用此规范，我们已经标注了约 3 万句汉语依存句法树库。

关键词：依存句法，标记规范

中图分类号：TP391 **文献标识码**：A

Annotation Guideline of Chinese Dependency Treebank from Multi-domain and Multi-source Texts

GUO Lijuan, LI Zhenghua, PENG Xue, ZHANG Min

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

Abstract: Dependency parsing has attracted much attention in the research community. There is no public, integrated and systematic annotation guideline for Chinese dependency treebank. Considering the special linguistic phenomena in web texts, this paper proposes a new annotation guideline for Chinese dependency treebank, which is adapted to multi-domain and multi-source texts. This annotation guideline aims to accurately depict the syntactic structures of various linguistic phenomena, and to ensure annotation consistency as well. Based on the proposed guideline, we have annotated about 30 000 Chinese sentences with their dependency structures.

Keywords: dependency; annotation guideline

0 引言

依存句法分析的目标是给定输入句子，构建一棵依存句法树，捕捉句子内部词语之间的修饰或搭配关系，从而刻画句子的句法和语义结构[1]。图 1 为一棵依存句法树的示例。其中，$ 表示一个伪词，指向句子根节点。作为依存树的最基本单元，一条依存弧包含三要素：核心词（父亲）、修饰词（儿子）和依存关系标签。例如，（我←有，subj）这条依存弧表示“有”为核心词，“我”为修饰词，依存关系标签为 subj（主语）。在此约定依存弧的方向由核心词指向修饰词。一棵合法的依存树必须满足两个条件：①单核心，即每个词只有一个核心词；②连通，即 $ 可沿弧的方向到达任何词。与短语结构句法相比，依存句法的优点是：①结构扁平，形式简单，容易理解，因此更适合普通人标注；②适用于不同语言；③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。因此，在过去十多年里依存句法分析得到越来越多的关注。

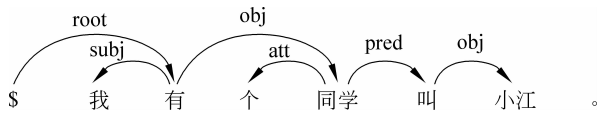

图 1 依存句法树示例

近几年来，随着深度学习在自然语言处理领域

收稿日期：2018-01-22 **定稿日期**：2018-03-13

基金项目：国家自然科学基金（61502325，61432013，61525205）

的快速发展，依存句法分析准确率也有了显著提高。以 CoNLL-2009 汉语标注评测数据集为例，基于传统离散特征的模型最好准确率（LAS）为 78.51%[2]。斯坦福大学 Chen 和 Manning 提出一个基本的利用前馈神经网络进行移进归约分类的依存句法分析方法，准确率为 77.29%[3]。Zhou 等[4]在 Chen 和 Manning 的方法中增加了全局正则化和概率优化，这一思路后来被谷歌采用并做了更好的网络优化，准确率达到 80.85% [2]。斯坦福大学 Dozat 和 Manning 提出在基于图的依存句法分析中，利用深层双线性神经网络进行依存弧分值预测，准确率达到惊人的 85.38%[5]。

虽然基于深度学习的依存句法分析方法在标准测试集上取得了 85.00%的准确率，然而众所周知，当处理有别于训练数据的文本时，依存句法分析的准确率会急剧下降。2012 年谷歌组织的 parsing the web 评测，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。评测结果发现，在英文新闻文本测试集上最高准确率为 91.88%，而在英文网络文本上的准确率只能达到 83.46%[6]。谷歌 2016 年基于神经网络的方法在英文新闻文本测试集上的准确率为 92.79%，在英文网络文本上的准确率为 87.54%，仍然有约 5%的差距[2]。

和英文相比，面向汉语网络文本的依存句法树构建进展更为迟缓，研究工作由于缺少一定规模的训练和评价数据而搁置。

基于上面的讨论，我们认为目前依存句法分析的最大挑战不是算法和模型的创新①，而是如何提高不同类型的网络文本上的依存句法分析准确率。考虑到自动领域移植方面的研究进展缓慢，我们认为最行之有效的方法就是数据标注。即对不同类型的网络文本，分别标注一定规模的训练和测试语料。

然而，现阶段依存句法树库的构建却存在很多的问题，主要体现在以下两个方面。

（1）目前学术界广泛使用的依存句法树库大部分是由短语结构树库基于规则自动转换而来。知名度很高的 Universal Dependency Treebank（UDT）②中包含了几十种语言的依存句法树库。然而，大部分语言都只有短语结构句法树库，需要通过基于规则的方法自动将短语结构转成依存结构，同时指定依存关系标签。经过仔细研究，我们认为 UDT 的依存句法结构和关系标签并没有考虑人工标注的需求，无法作为一个严格的标注规范指导人工进行高质量的标注。例如，在 45 种依存关系标签中，存在一些实际标注中很难区分的关系标签。

（2）目前还没有一个公开、完整、系统的汉语依存句法树标注规范。哈尔滨工业大学在 Linguistic Data Consortium（LDC）③上发布了一个 5 万句的汉语依存句法树库，本文称之为 Harbin Institute Technology Chinese Dependency Treebank（HIT-CDT）[7]。邱立坤、金澎等标注了一个大规模的汉语依存句法树库，但是目前还没有公开发布这个数据，本文称之为 Peking University Chinese Dependency Treebank（PKU-CDT）。同时，他们对 HIT-CDT 标注规范进行了扩充，以便将依存结构转为短语结构[8]。然而，这两个树库并没有公开发布一个完整、系统的标注规范。

本文提出了一个新的适应多领域多来源文本的汉语依存句法数据标注规范。按照此规范，我们已经标注了约 3 万句依存句法树库，并将其命名为“Soochow University Chinese Dependency Treebank（SU-CDT）”。最新的标注规范（不断更新）和最新树库（不断扩大）我们将发布在 http://hlt.suda.edu.cn/index.php/SUCDT。

1　编制标注规范的考虑因素

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。为了达到这个目标，我们必须制定一个科学（满足语言学理论）、系统（条理清晰、容易掌握）、完整（覆盖各种语言现象）的标注规范，作为整个工作的基础，从而提高不同标注者之间的一致性，保证标注质量。本文第二作者于 2010 年夏，主持了哈工大 HIT-CDT 树库的整个标注过程。本文提出的标注规范充分借鉴了哈工大标注规范，同时吸取了 HIT-CDT 标注过程中的经验教训。在此，标注规范编制的初衷和考虑因素总结如下：

（1）针对汉语，设计一个尽可能精简的依存关系标签集合。依存关系标签数量过多，会大大增加标注难度。例如，目前 UDT 的依存关系标签有 40 多种。但是我们仔细研究后认为，UDT 中关系标签

① 如何将语义知识，如动词和名词之间的配价关系，加入到深度学习模型中，也是很有挑战且非常有价值的研究方向。

② http://universaldependencies.org/treebanks/zh.cfl/index.html

③ https://catalog.ldc.upenn.edu/LDC2012T05

存在两个问题：①主要面向英语等印欧语系语言设计；②关系标签分类过细，实际标注时区分难度很大。哈工大标注规范一共只有 14 种依存关系标签，我们进一步精简。例如，哈工大规范中左附加(LAD)和右附加(RAD)，只是根据依存弧的方向区分，因此合并为一个附加关系(adjct)。

(2) 设计一个完整的依存关系标签集合，充分刻画汉语的不同语言现象。我们在哈工大标注规范的基础上，增加了一些关系，如 app(称呼)、exp(进一步解释)、frag(片段)等，以刻画不同语言现象(口语化、不规范表达，甚至病句)。目前我们的规范包含了 20 种依存关系标签，如表 1 所示。

表 1 依存关系标签汇总表

关系标签	说明	例句	标注结果
root	sentence root(根节点)	我 爱 妈妈	$\xrightarrow{root}$爱
sasubj-obj	same subject and object(同主语同宾语)	图 3 例句	建立$\xrightarrow{sasubj\text{-}obj}$健全
sasubj	same subject(同主语)	图 3 例句	建立$\xrightarrow{sasubj}$改进
dfsubj	different subject(不同主语)	图 3 例句	建立$\xrightarrow{dfsubj}$提高
subj	subject(主语)	我 爱 妈妈	我$\xleftarrow{subj}$爱
subj-in	subject inside a subject-predicate predicate(主谓谓语中的内部主语)	他 确实 头 疼	头$\xleftarrow{subj\text{-}in}$疼
obj	object(宾语)	我 爱 妈妈	爱$\xrightarrow{obj}$妈妈
pred	predicate(谓语)	命令 他 扫地	他$\xrightarrow{pred}$扫地
att	attribute modifier(定语)	国家 主席	国家$\xleftarrow{att}$主席
adv	adverbial modifier(状语)	非常 喜欢	非常$\xleftarrow{adv}$喜欢
cmp	complement modifier(补语)	洗 干净 手	洗$\xrightarrow{cmp}$干净
coo	coordination construction(并列结构)	鲜花 和 掌声	鲜花$\xrightarrow{coo}$掌声
pobj	preposition object(介宾)	在 家 看书	在$\xrightarrow{pobj}$家
iobj	indirect-object(间宾)	给 他 书	给$\xrightarrow{iobj}$他
de	de-construction(“的”字结构)	这 是 他 的	他$\xleftarrow{de}$的
adjct	adjunct(附加成分)	我 走 了	走$\xrightarrow{adjct}$了
app	appellation(称呼)	老师 ，你 好	老师$\xleftarrow{app}$好
exp	explanation(进一步解释)	普京(俄罗斯 总统)	普京$\xrightarrow{exp}$总统
punc	punctuation(标点)	我 爱 妈妈。	爱$\xrightarrow{punc}$。
frag	fragment(片段)	你，我，中国。	你$\xrightarrow{frag}$我$\xrightarrow{frag}$中国

(3) 以谓语为核心，尽可能丰富地刻画复杂句子内部结构。哈工大规范使用一个独立结构关系(IS)来标注并列谓语之间的关系。我们将其细化为 sasubj(同主语)、sasubj-obj(同主同宾)和 dfsubj(不同主语)三种依存关系标签，从而更深入地表示句子内多个谓语之间的关系，并为上层语义分析提供支持。

(4) 适应不同分词粒度。由于汉语中由词素组成词，由词组成短语时，界线很模糊，因此学术界对于分词的粒度没有一个统一的界定。我们在规范制定过程中充分考虑了这一因素，并给出一些不同分词粒度下的标注示例。例如，“走向世界”中，如果“走向”作为一个词，那么“世界”为宾语；如果作为两个词，那么“世界”是“向”的介词宾语，“向”作为补语修饰“走”。

(5) 尽可能准确地刻画语义结构。在满足规范

中阐述的具体规则的前提下，选择最能准确表达语义关系的依存树。如图 2 所示，“预计”的主语省略，而不是“教学楼”，将“教学楼明年竣工”标注成“预计”的宾语从句，这样才最能准确表达语义。这种存在交叉弧的依存树又称为非投影树。我们发现，由于汉语语序灵活，一小部分句子的确需要用非投影树标注。

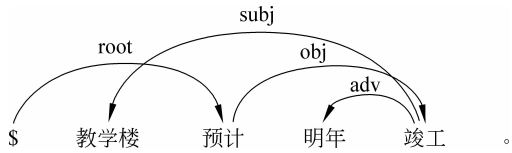

图 2　非投影树示例

（6）当两种标注都满足规范，且符合语义时，我们一定会明确规定一个优先顺序，帮助标注者选择，从而有效提高标注一致性。

随着标注实践的进行，我们会深入研究实际标注中遇到的问题，积极与其他专家交流，不断学习语言学论著，持续完善和更新我们的规范。

2　依存关系标签详细介绍

此规范目前包含了 20 种关系标签，分为三个大类：

（1）谓语对应的依存关系标签：root，sasubj-obj，sasubj 和 dfsubj，此类依存关系标签全部为右弧，用于标注主要谓语关系；

（2）单句内部主干关系标签：subj，subj-in，obj，pred，att，adv 和 cmp，用于标注汉语句子中的主谓宾定状补关系结构；

（3）单句内部其他关系标签：coo，pobj，iobj，de，adjct，app，exp，punc 和 frag，用于辅助标注汉语句子的其他关系结构。

以下将对这些依存关系标签逐一展开介绍。

2.1　谓语对应的依存关系标签

汉语中谓语是用于说明或陈述主语的动作或状态。动词、形容词、名词、介词、主谓结构等都可以充当谓语。

root（sentence root，根节点）：规定句子的第一个主要谓语以 root 关系修饰伪节点 $。因为句子是可以嵌套的，即可以有主语从句、宾语从句、定语从句等，所谓“主要谓语”是指句子最顶层的一个或多个谓语，而不是在从句中的谓语。

sasubj-obj（same subject and object，同主语同宾语）：规定当两个同级的谓语共享主语和宾语时，后一个谓语以 sasubj-obj 关系修饰前一个谓语。

sasubj（same subject，同主语）：规定当两个同级的谓语共享主语但不共享宾语时，后一个谓语以 sasubj 关系修饰前一个谓语。

dfsubj（different subject，不同主语）：规定当两个同级的谓语具有不同主语时，后一个谓语以 dfsubj 关系修饰前一个谓语。

2.2　单句内部主干关系标签

用于标注主谓宾定状补结构。

subj（subject，主语）：主语是谓语的描述对象、施事或受事。由于这三种情况属于语义的范畴，并且区分起来对标注者的要求过高，因此大多数情况下我们不对其进行详细区分，具体介绍见第 4 节中对主语和宾语的明确规定。

subj-in（subject inside a subject-predicate predicate，主谓谓语中的内部主语）：句子中一个主谓短语整体作为谓语，称为主谓谓语[9]。以 subj-in 专门标注主谓谓语的内部主语。

obj（object，宾语）：和主语类似，宾语是谓语的受事或施事，但通常位于谓语的后面。

pred（predicate，谓语）：用来刻画汉语中独特的兼语结构，如图 1 所示。

att（attribute modifier，定语）：定语是名词或代词的修饰成分，通常位于核心词的前面。

adv（adverbial modifier，状语）：状语是动词或形容词的修饰成分，通常位于核心词的前面。

cmp（complement modifier，补语）：补语是动词或形容词的修饰成分，通常位于核心词的后面。

2.3　单句内部其他关系标签

coo（coordination construction，并列）：多个句法功能相同的词（非谓语）并列在一起，通常中间会用“和”“与”或顿号连接，我们规定后一个词以 coo 关系修饰前一个词，形成波浪状。

pobj（preposition object，介宾）：介词和宾语构成介宾短语时，宾语用 pobj 关系修饰介词。

iobj（indirect-object，间宾）：“给/送/授予/称呼/叫”等动词后面可以跟两个名词性宾语，为了区分，第一个宾语称为间接宾语，以 iobj 关系修饰动词。

de（de-construction，“的”字结构）：“的”字后面

的名词或代词明显省略的情况，例如，“我喜欢红色的”，修饰词“红色”以 de 关系修饰“的”。

adjct(adjunct，附加成分)：句子中没有实际意义的、只是为了让句子结构完整、或者讲起来更有韵味(抑扬顿挫)的词，统一标注为附加关系。

app(appellation，称呼)：口语中句子最前面对人的称呼语，以 app 关系修饰句子第一个主要谓语。

exp(explanation，进一步解释)：汉语中常用括号中的内容或者冒号后面的内容，对前面的词、短语或句子进一步解释说明，规定解释性的内容以 exp 关系修饰被解释的内容。

punc(punctuation，标点)：规定标点以 punc 关系修饰核心词。

frag(fragment，片段)：网络文本中出现的不符合语法、支离破碎的病句，后一个成分以 frag 关系修饰前一个成分，形成波浪状。

3 标注规范的几点重要创新

同主语关系：从句法的角度看，谓词是句子中最重要的词。和英语不同，汉语中可以使用标点符号直接将几个谓语句连成一个句子。如何确定多个谓语之间的搭配关系，是标注规范必须妥善回答的问题。哈工大规范使用独立结构关系(IS)和并列关系(COO)来标注多个谓语之间的关系，然而实际标注中很难把握其界线，标注者甚至需要考虑多个谓语句之间的语义逻辑关系，导致很多分歧。

根据目前学术界的标准，多个谓语词之间的逻辑语义关系，属于语义和篇章分析的范畴，因此一般作为语义和篇章分析任务的处理对象。我们的规范的主要目标是：在保证标注一致性和质量的前提下，充分刻画句子的句法结构。而多个谓语之间的逻辑语义关系确实太复杂了，因此我们的规范明确规定不考虑多个谓语句之间的语义逻辑关系，仅仅考虑句法关系。根据多个谓语是否共享主语和宾语，细分出三种依存关系标签：sasubj、sasubj-obj 和 dfsubj，以便更深入地表示多个谓语之间的关系。这样不仅可以标注出句子的谓语信息，同时为上层语义标注和分析提供支持，示例如图 3 所示。

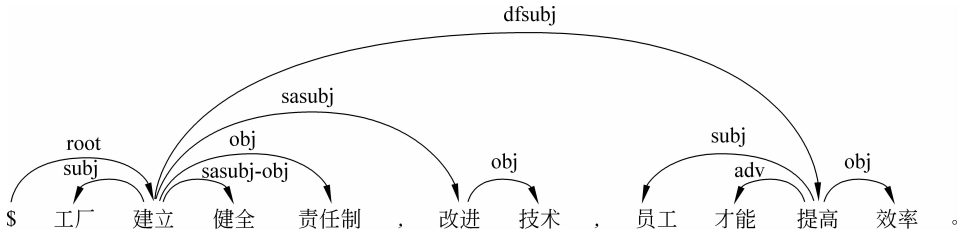

图 3 同主语关系示例

我们规划未来在目前句法结构的基础上，进一步标注语义信息，即作为一个新的语义标注任务，制定规范，进而人工标注。而不是尝试在一个任务下把所有的信息都标注上。

内部主语关系：赵元任先生提出“汉语句子里主语和谓语的语法意义是主题(topic)和述语(comment)，而不是施动者(actor)和动作(action)”的观点[10]，引发了汉语研究的新一轮思考。石定栩先生的文章中讲到“大部分语言学家主张主题和主语都是汉语句子的成分，而且具有不同的句法地位(省略原文中的引用)。不过，对于主题和主语的定义及其在句法过程中的地位，则还没有定论。常见的做法之一是将主题或主语的功能和分布情况一一列举。然后以这些功能和分布作为标准，判断某一成分是主语还是主题。”[11]

从以上讨论可以看出，汉语中主题和主语的区分，是非常困难的事情，需要很强的语言学专业背景和细腻的语感。哈工大标注规范采用了回避和简化的策略，将主题也当作主语，允许一个谓语具有多个主语，如图 4 所示。我们延续哈工大标注规范的策略，不区分主题和主语，从而保证标注者的一致性。

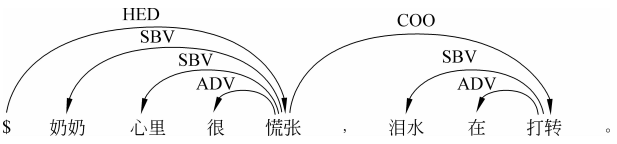

图 4 哈工大规范双主语标注示例

按照哈工大规范，“慌张”对应两个主语“奶奶”和“心里”，并且将谓语“慌张”和“打转”以 COO 标注为并列。

俞士汶先生等在其《现代汉语语法信息词典详解》前言(第 2 版)中提到：“在主谓结构中，不仅主语可以由另一个主谓结构来充任，而且谓语也可以

由另一个主谓结构来充任(这就形成了所谓的‘主谓谓语句’或‘主谓谓语短语’)。”[9] 这一观点在朱德熙先生的《语法讲义》[12] 中也得到印证。受这种观点的启发，我们提出 subj-in 这个依存关系标签，专门标注主谓谓语的内部主语。虽然我们不刻意区分主题和主语，但在很多情况下，subj-in 可以标注出主谓谓语的内部主语(一般是谓语词的主语)，而 subj 可以标注出主谓结构的主语(一般是句子的主题)。如图 5 所示，“奶奶”实际上是整个句子的主题，同时也是主谓谓语“心里很慌张”和“泪水在打转”的主语；“心里”是“很慌张”这个谓语词的主语；“泪水”是“打转”这个谓语词的主语。对比图 4 和图 5，我们认为 subj-in 带来几点优势：①将主谓谓语动词的主语和句子的主题区分开；②方便刻画多个谓语的同主语关系；③体现主谓谓语整体作为一个组块的信息(传统依存结构实际上没有组块信息)。

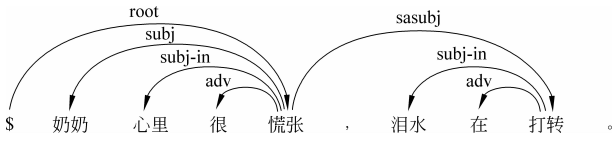

图 5　苏大规范双主语标注示例

按照我们的规范，“奶奶”是主谓谓语“心里很慌”和“泪水在打转”的主语，并且以 sasubj 将两个主谓谓语标注为同主语关系；“心里”是“很慌张”这个谓语词的主语；“泪水”是“打转”这个谓语词的主语。

自从增加了 subj-in 这个依存关系标签后，我们标注实践时发现，符合“N_1＋N_2＋谓语”结构的句子(其中 N_1 和 N_2 分别表示两个名词)，通常都适合标注为含有 subj-in 结构(N_1←谓语，subj；N_2←谓语，subj-in)，因此从一定程度上验证了主述位理论的合理性。然而，这样的句子同样也可以按照传统的 att 结构标注(N_1←，att；N_2←谓语，subj)，两种标注之间的界线很难区分，我们既不能规定全部标注为 att，也不能全部标注为 subj-in。为了提高标注结果一致性，我们明确规定当两种标注方法都适用时，只有下面两种情况标注为 subj-in ：

(1) 标注为 subj-in，可以进一步捕获同主语关系，如图 5 所示；

(2) 标注为 subj-in，可以避免交叉弧的出现，如图 6 所示。

除以上两种情况外，我们的规范将这种结构标注成(N_1←，att；N_2←谓语，subj)

对主语和宾语的明确规定：我们了解到，从语言学角度上看，“施事”和“受事”属于语义范畴，而主语和宾语属于句法层面，并且目前句法分析标注规范通常都只标注到句法层面。我们在哈工大依存树库标注中发现，对主语进行语义上的细分有时候非常困难，很难给出一个统一的标准。例如“经济发展得很快”这个句子中，有的标注者认为“经济”是“发展”的对象(即受事)，有的标注者则认为“发展”是对“经济”的状态的描述(即描述对象)，标注一致性很低。因此我们要求标注者根据焦点词和谓语的相对位置，选择 subj 或 obj。也就是说，在实际标注过程中，标注人员大多数情况下不用区分“施事”和“受事”，直接将谓语前面的作为主语，谓语后面的作为宾语，如图 7 所示。

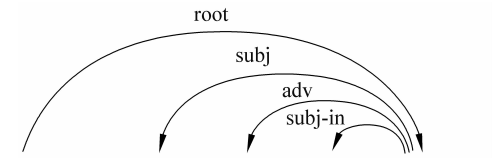

图 6　标注 subj-in 避免交叉弧示例

若标注为(公司←效益，att ；确实←不佳，adv)，会产生交叉，所以标注为 subj-in。

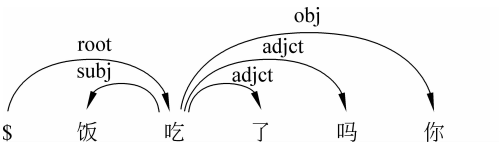

图 7　对主语和宾语的明确确定示例 1

当“施事”和“受事”同时在谓语的左边或右边出现时，为了避免一个谓语有两个主语或两个宾语，我们要求标注者严格区分“施事”和“受事”，将“施事”标为主语，“受事”标为宾语，如图 8 所示。我们发现，这种情况在实际标注中遇到的概率非常低，并且标注者很容易区分“施事”和“受事”，歧义很小，一致性很高。

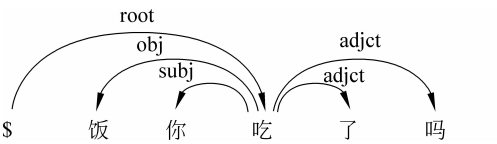

图 8　对主语和宾语的明确确定示例 2

兼语结构的标注：兼语结构(V_1＋N＋V_2)是汉语中的一种非常独特的语法结构，其中 N 是 V_1 的宾语，同时又是 V_2 的主语。为了准确表达这种结构，我们引入 pred 这个关系标签，打破主语修饰

谓语的惯例，让谓语 V_2 直接修饰主语 N，如图 1 所示（有→同学，obj；同学→叫，pred）。和 HIT-CDT 中（有→同学，DBL；有→叫，VOB）的标注形式相比，我们认为 pred 的引入，让语义上更为紧密的 N 和 V_2 直接连接，因此是一种更好的表达形式。

复合名词短语内部结构的标注：汉语中有很多形如“$W_1W_2W_3$”的复合名词短语，名词“W_3”是整个短语的核心词，难点在于其内部的结构如何标注，即需要确定（W_1←W_2，att）或（W_1←W_2W_3，att），这里可以把“W_2W_3”看成一个词。我们的规范首次明确规定了复合名词内部标注的优先级规则：仔细分析内部的语义搭配强度（W_1←W_2W_3，att）vs.（W_1←W_2，att）；如果两个标注强度没有明显的差别，则优先标注成（W_1←W_2W_3，att），如图 9 所示；如果两个标注强度有明显的差别，则按照标注强度标注，如图 10 所示。

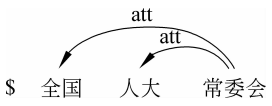

图 9 复合名词短语内部标注示例 1

（全国←常委会，att；人大←常委会，att）的修饰强度和（全国←人大，att；人大←常委会，att）没有明显的差别，所以规定标注为前者。

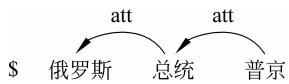

图 10 复合名词短语内部标注示例 2

（俄罗斯←总统，att；总统←普京，att）的修饰强度要高于（俄罗斯←总统，att；俄罗斯←普京，att），所以规定标注为前者。

4 汉语依存句法树标注规范的标注实践

为了更好地支持依存句法分析树的标注，我们自 2014 年起开发了一个标注系统，并根据实际需求不断完善。此标注系统中主要设计了三种核心角色：①标注人员，标注分配的任务，也可以对专家的答案提出投诉；②审核专家，对两个标注人员标注不一致的任务进行审核，并确定唯一答案。需要注意的是，同一个标注任务的两个标注结果中只要有一条依存弧不相同，就会触发审核。审核界面中会把不相同的地方突出出来，以方便标注人员对比；③高级专家处理标注人员的投诉任务，确定最终答案。

图 11 给出了一个任务（句子）的标注流程。首先，标注系统会将一个任务随机分配给两个标注人员标注。标注完成后，如果两个标注结果完全一致，那么就认为已确定答案，流程结束。如果两个标注结果至少有一条弧不一致，就会触发审核机制，系统会将这个任务随机分配给一位专家进行审核，确定唯一答案。进而，标注系统会将审核过的答案，反馈给出错的标注人员进行学习。学习过程中，如果标注人员对答案不认可，可以提出投诉。如果没有出现投诉，那么就认为已确定答案，流程结束。如果出现投诉，系统会将投诉任务随机分配给一位高级专家，确定唯一答案。标注人员投诉、审核专家审核及高级专家处理投诉时，可以把各自的理由写出来，从而实现非常有效的异步沟通。除此之外，我们还会在线下通过在线聊天工具就一些问题进行交流、搜集反馈、修改答案、完善规范。

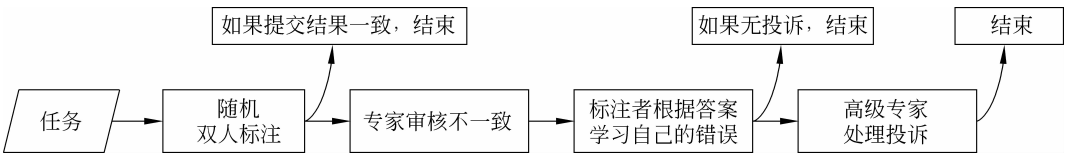

图 11 标注流程图

为了最大程度降低人工付出，一方面我们开发了一个基于浏览器的在线标注系统，减轻了数据标注管理者的负担；另一方面对于选取的数据，我们采用局部标注的方式，即只选择句子中模型置信度较低的一定比例的词，进行标注[13]，从而节省标注时间和成本，并且增强标注者的注意力。同时，在一批新的数据批次中我们会将以前标注过的有答案的任务作为地雷混入，我们称之为地雷机制。我们放入地雷有两大作用：①自动评价标注人员的标注情况；②进一步检查之前的标注结果，以便提高标注答案质量。标注的过程中，随着规范的更新，也需要更新以前的标注结果。

总之，我们希望标注系统设计和标注流程管理处处从提高质量的目标出发，并且最大程度减少数

据标注管理者的工作，将数据管理尽可能科学化、系统化，为大规模数据标注提供便利。

为了持续标注大规模的依存句法数据，我们组织了几十位苏州大学本科生作为兼职数据标注人员。首先我们向标注人员详细介绍我们的规范以及标注系统的使用。进而，标注人员系统学习标注规范，并且在标注系统上模拟训练。最后，标注人员进行真实数据标注工作。经过一定时间考察，我们会选择标注质量高的标注人员作为审核专家。到目前为止我们通过标注系统共标注了约 3 万句依存句法数据，数据的来源见表 2。

表 2　数据来源说明表

来源	领域	句子数
哈工大 CDT	《人民日报》、小学课本	约 1 万句
PCTB7	新闻（杂志、广播）、广播对话、讨论组、博客	约 1 万句
阿里内容搜数据	淘宝头条	约 1 万句

我们通过对句子的标注结果进行统计与分析发现，和最终答案相比，标注者的平均依存弧准确率为 87.6%，标注者之间的平均依存弧一致率为 76.5%。而标注者之间平均句子级别的一致率只有 43.7%，即 56.3% 的句子需要审核专家进一步检查。这表明了句法标注工作的困难性，以及为了保证标注质量，需要严格双人标注的重要性。

5　结语与展望

本文提出了一个新的适应多领域多来源文本的汉语依存句法数据标注规范，以指导大规模实际标注工作。该规范考虑了多方面的因素，同时参考一些经典的语言学著作，设计了 20 个依存关系标签，适应于多领域多来源文本的汉语依存句法数据标注，且可以尽可能准确地刻画大部分汉语文本的句子级句法结构；同时，该规范对很多难以理解并区分的语言现象进行了比较详细的总结。实际标注结果表明，根据我们的标注规范，可以达到较高的标注一致性。

未来我们会按照该规范持续标注多领域多来源文本，提高依存句法分析准确率，也为领域移植研究工作提供数据支持。同时，我们会总结实际标注过程中遇到的问题，不断完善和更新。目前的规范可以满足表 2 中数据的标注需求，但是未来如果遇到规范不能涵盖的语言现象，我们会增加新的依存关系标签，扩充我们的规范。

参考文献

[1] 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.

[2] Andor D, Alberti C, Weiss D, et al. Globally normalized transition-based neural networks[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 2442-2452.

[3] Chen D, Manning C. A fast and accurate dependency parser using neural networks[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 740-750.

[4] Zhou D, Zhang Y, Huang S, et al. A neural probabilistic structured-prediction model for transition-based dependency parsing[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015: 1213-1222.

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

[6] Petrov S, Google R M, York N, et al. Overview of the 2012 shared task on parsing the web[C]//Proceedings of the Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL), 2012.

[7] Liu T, Ma J, Li S. Building a dependency treebank for improving Chinese parser[J]. Journal of Chinese Language and Computing, s 2006(16): 207-224.

[8] 邱立坤，金澎，王厚峰. 基于依存语法构建多视图汉语树库[J]. 中文信息学报，2015，29(3)：9-15.

[9] 俞士汶，等. 现代汉语语法信息词典详解[M]. 北京：清华大学出版社，2003.

[10] Chao Y. A grammar of spoken Chinese[M]. London: University of California Press, 1968: 69.

[11] 石定栩. 汉语主题句的特性[J]. 现代汉语，1998(2)：44-59.

[12] 朱德熙. 语法讲义[M]. 北京：商务印书馆，1981：106.

（下转第 52 页）

Massachusetts,Persus Publishing,2002.
[11] 刘知远,孙茂松.汉语词同现网络的小世界效应和无标度特性[J].中文信息学报,2007,21(6):52-58.
[12] 刘知远,郑亚斌,孙茂松,等.汉语依存句法网络的复杂网络性质[J].复杂系统与复杂性科学,2008,5(2):37-45.
[13] Cancho R F I,Sole R V. The Small World of Human Language[J]. Proceedings of the Royal Society of London Series B-Biological Sciences, 2001, 268(1482): 2261 -2265.
[14] Cancho R F I, Sole R V, Kohler R. Patterns in Syntactic Dependency Networks[J]. Phys Rev E,2004,69(5): 051915.
[15] Motter A E,de Moura A P, Lai Y C,et al. Topology of the conceptual network of language[J]. Phys Rev E, 2002,65(6): 065102.
[16] 青海省科技查新检索咨询中心.基于复杂网络的藏文基本单位统计特征研究[R]. QH-10090-N, 2016-12-21.
[17] Erdos P,Renyi A. On the evolution of random graphs [J]. Publ Math Inst Acad Sci,1960,(5): 17-61.
[18] Watts D J. The new science of networks[J]. Annual Review of Sociology,2004(30): 243-270.
[19] Newman MEJ. The structure and function of complex networks[J],SIAM Rev,2003,45(2): 167-256.
[20] Watts D J, Strogatz S H. Collective dynamics of 'small-world' networks[J]. Nature, 1998, 393: 440-442.
[21] Barabasi A L, Albert R. Emergence of scaling in random networks[J]. Science,1999,286 (5439): 509-512.
[22] Liang W, Shi Y, Tse C K, et al. Comparison of co-occurrence networks of the Chinese and English languages[J]. Physica A, 2009,388: 4901-4909.
[23] 才智杰.藏文自动分词系统中紧缩词的识别[J].中文信息学报,2009,23(1): 35-37.
[24] 郭进利,汪丽娜.幂律指数在 1 与 3 之间的一类无标度网络[J]. 物理学报,2007,56(10): 5635-5639.
[25] 王林,戴冠中.复杂网络的度分布研究[J].西北工业大学学报,2006,24(4): 405-409.

才智杰(1970－),博士研究生,教授,主要研究领域为藏文信息处理、藏语自然语言处理。
E-mail: Czjqhsd@163.com

孙茂松(1962－),博士,教授,博士生导师,主要研究领域为自然语言处理和人工智能。
E-mail: sms@tsinghua.edu.cn

才让卓玛(1970－),博士,教授,主要研究领域为人机语音交互、藏文信息处理。
E-mail: cr-zhuoma@163.com

(上接第 35 页)

[13] Li Z, Zhang M, Zhang Y, et al. Active Learningfor Dependency Parsing with Partial Annotation[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 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: lj_guo0113@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

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