基于篇章的汉语句法结构树库

卢 露 [1] 矫红岩 [1] 李 梦 [1] 荀恩东 [1]

摘 要 为快速构建一个大规模、多领域的高质树库, 提出一种基于短语功能与句法角色组块的、便于标注多层次结构的标注体系, 在篇章中综合利用标点、句法结构、表述功能作为句边界判断标准, 确立合理的句边界与层次; 在句子中以组块的句法功能为主, 参考篇章功能、人际功能, 以 4 个性质标记、8 个功能标记、4 个句标记来描写句中 3 类 5 种组块, 标注基本句型骨架, 突出中心词信息. 目前已初步构建有质量保证的千万汉字规模的浅层结构分析树, 包含 60 余万小句的 9 千余条句型结构库, 语料涉及百科、新闻、专利等应用领域文本 1 万余篇; 同时, 也探索了高效的标注众包管理模式.

关键词 语料库标注, 树库, 语块, 句法分析

引用格式 卢露, 矫红岩, 李梦, 荀恩东. 基于篇章的汉语句法结构树库. 自动化学报, 2022, **48**(12): 2911−2921

DOI 10.16383/j.aas.c190828

A Discourse-based Chinese Chunkbank

LU Lu[1] JIAO Hong-Yan[1] LI Meng[1] XUN En-Dong[1]

Abstract In order to provide a large scale annotation of Chinese functional chunk for linguistic research and syntactic parsing, we present a method to quickly build a discourse based Chinese chunkbank with high quality in multi-domain: Firstly, we use punctuations, syntax, expression functions of VP and NP, to segment complex sentences into several independent simple sentences; Secondly, based on the syntactic function, textual function, discourse function and interpersonal function of the chunks, we design 4 phrase tags, 8 functional tags, 4 sentence boundary tags to depict the chunks, which was classified into 3 types and 5 kinds. the annotators annotated the skeleton structure and highlighted the head word of the predicate for every simple sentence. Until now, we have been annotating more than 10 million of Chinese characters, including 9 thousand of skeleton structures for 60 thousand sentences. The chunkbank covers a range of text genres, including baidubaike, internet news, patent, etc. At the same time, we explored an effective model of crowdsourced data management.

Key words Corpus annotation, treebank, chunk, syntactic parsing

Citation Lu Lu, Jiao Hong-Yan, Li Meng, Xun En-Dong. A discourse-based Chinese chunkbank. *Acta Automatica Sinica*, 2022, **48**(12): 2911−2921

20 世纪 90 年代以来, 汉语开发了很多体系成熟、影响较大、规模不等的树库, 短语结构树以宾州中文树库 (Chinese treebank, CTB)[1]、北京大学汉语树库 (Peking University treebank, PCT)[2]、清华汉语树库 (Tshinghua Chinese treebank, TCT)[3] 等为代表, 依存结构树主要有 Chen 等[4] 的 Sinica、哈尔滨工业大学中文依存树库 (HIT Chinese dependency treebank, HIT-CDT)[5]、苏州大学汉语依存树库[6] 等. 短语结构树细致地描写了句子结构层次、短语的类别与功能, 转换为浅层结构树与依存关系树较为容易, 但往往中心词、语义关系不突出, 节点与标签较多, 计算开销大; 依存结构树突出中心词信息及依存关系, 便于转化为语义依存描述[7], 同时, 计算开销较少, 符合语言直觉而更易于标注, 但缺乏短语类别与整体功能信息, 缺乏明确依存关系的现象较多, 长距离依存也难以被解析. 一些树库尝试融合短语结构树与依存树的优点, 如 TCT、北京大学多视图依存树库 (Peking University multi-view Chinese treebank, PMT)[8]. 为克服长句分析不理想的问题, 组块分析采取 “分而治之”的策略, 北京大学中文语块库 (Chinese chunkbank)[9] 根据句法功能, 无嵌套地标注句子骨架, 中文命题树库 (Chinese proposition bank, CPB)[10] 探索 “谓词−论元”结构 (Predicate-argument structure) 的语义组块分析, 这些树库初步探索了汉语浅层分析树库构建、组块自动分析; 随着篇章级句分析及句间关

收稿日期 2019-12-05 录用日期 2020-04-10
Manuscript received December 5, 2019; accepted April 10, 2020
国家社会科学基金 (16AYY007), 北京语言大学研究生创新基金 (19YCX121) 资助
Supported by National Social Science Fundation of China (16AYY007) and Graduate Research and Innovations Fundation of Beijing Language and Culture University (19YCX121)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 北京语言大学信息科学学院 北京 100083
1. College of Information Science, Beijing Language and Culture University, Beijing 100083

系日渐被重视, TCT 与 PCT 则在构建树库的同时, 以传统单复句理论为指导, 区分单复句、句群, TCT 同时标注句间关系, PCT 则初步区分了篇章标记; 而一些篇章理论与语料库则不同程度地探讨了显明篇章标记的作用与分布, 并结合实际需求深入探讨了汉语 “句”边界问题[11].

整体而言, 现有树库在规模与句法表示方面主要面临两个挑战: 1) 大规模标注困难. 全树句法分析技术在大规模真实语料上正确率不高[12], 现有树库标注体系及加工模式, 人工扩大规模难以既保证标注质量, 又兼顾标注规模与速度. 2) 对虽在句中但属于篇章层面、语用表达层面的成分处理牵强, 要么将其纳入句法分析范畴, 增加了句法分析的冗余度与难度, 要么被看作一种可忽略或消除的噪音[1−2], 忽视了篇章的完整性与衔接连贯性, 不便于树库系统地扩展标注层次, 也不利于分析非标准书面语文本.

此外, 大部分树库主要采用 20 世纪末至 21 世纪初的新闻杂志语料, 领域相关性问题促使不少树库, 加入了部分学术科技类、应用类、口语类等非标准书面语语料[1, 3, 6, 8, 13]. 然而, 一方面, 句法结构标注难度限制了标注速度与规模, 各树库主要在 100 ~ 250 万汉字之间; 另一方面, 忽视了大规模真实语料中标点往往具有高歧义性、缺失性、错误性的现象, 很难在标点断句的基础上实现有效句法分析及标注的高一致性.

本文提出一种利于大规模加工的浅层句法标注体系: 以明确的断句方法, 尽可能地反映开放域篇章中句子的结构, 提高标注一致性; 根据短语功能与句法角色, 将句子分析为由句法成分、衔接成分、辅助成分构成的块状组合序列, 以组块状短语结构树为句法表示 (见图 1), 直接根据各组块的性质及功能, 标注句子骨架, 突出中心词信息. 以该标注体系为指导, 初步完成树库 1.0 构建工作: 包含 Kappa 值大于 0.8 的合格文本 1 万余篇, 共计 1 千余万字, 由于树库构建将以短语结构标注为基础, 分级分层逐步完成缺省结构、句间结构标注, 因此先行构建浅层句法结构树库, 因此为后续应用任务及组块依存结构标注、句间关系标注奠定了基础。

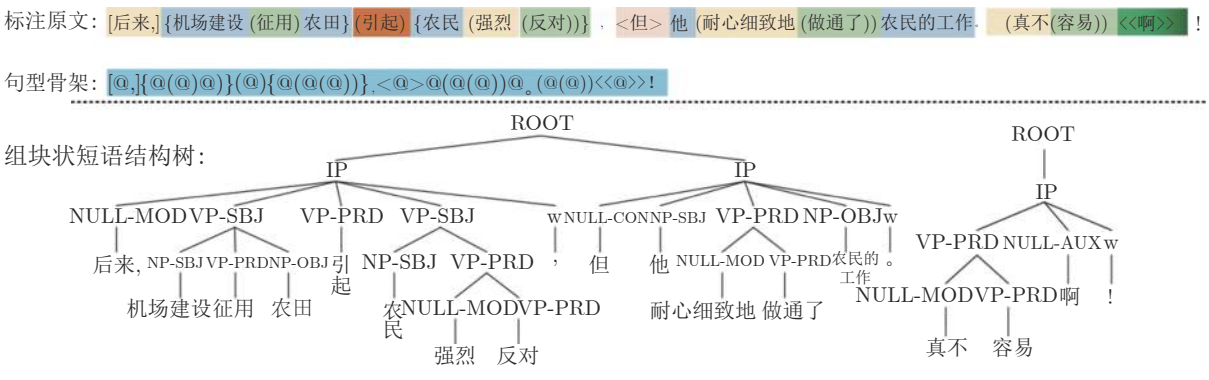
本文结构如下: 第 1 节介绍本树库的设计, 第 2 节介绍本树库标注实践, 第 3 节对树库已有数据进行统计分析, 第 4 节为总结和展望.

1 基于篇章的短语结构树库设计

尽可能保持树库中篇章语料完整性, 不作删减; 句子成分, 不独以句法功能作为划分标准及取舍标准, 也考虑部分成分表达句间语义结构关系的篇章功能[11]、表达语气的复杂功能[14], 对这些成分不勉强纳入句法分析, 也不笼统归入可忽略部分, 根据功能与用途加以分析; 从而逐步构建一个以短语结构标注为基础, 包含组块依存及话题共享结构、句间关系结构标注的多层次树库, 本文称这种树库为基于篇章的树库. 本文主要介绍先行构建的、以浅层骨架分析为主的短语结构树库.

1.1 组块状短语结构树标注设计

在篇章中分析句子, 无论是自底向上, 还是自顶向下, 都绕不开分句, 建立合理又系统的标准以确立小句边界是首要任务, 继而才能在小句的基础上探讨组块边界与结构, 最后根据所确定的语块关系设计标注符号; 为说明方便, 本文先假定已进行了分句处理, 从组块及标注符号开始说明, 最后介绍句边界标准设定 (详见第 2.2.3 节).



后来, 机场建设征用农田引起农民强烈反对, 但他耐心细致地做通了农民的工作。真不容易啊!

图 1 组块状标注结果及浅层分析的树结构示例

Fig. 1 A sample of shallow chunk-based syntactic tree

1.1.1 短语结构树中组块的种类与定义

篇章中的句子之间并非孤立存在, 而是通过指称、结构衔接、逻辑连接等手段组织起来的[15], 因此, 一些句子成分作为组织篇章的手段, 虽不一定参与句法构造, 但对篇章分析尤为重要, 不能将其笼统地看作冗余成分; 而汉语中一些表达语气、态度、意图等虚词也并非全然是句法分析层面的问题. 为尽可能清楚地描写 “谓词–论元”结构为中心的句子骨架, 本文在句中根据词或短语在句中的功能与用途, 将组块分为 3 类: 构成基本句子结构的句法成分组块; 起衔接上下文的衔接组块; 表达附加性语义的辅助组块.

短语结构标注的主要任务是解决句子结构层次问题, 但同时也尽可能为后续组块依存与话题结构、句间结构标注打下基础, 树库以充当谓语的最长短语组块中的谓词为核心, 标注最长主语块、宾语块与谓词的主谓宾结构, 以及最长状语块、补语块与谓词的状中补结构, 目的在于识别如下功能块边界及相互之间关系:

充当谓语的最长短语块— “述语”, 述语包括一个核心谓词, 及修饰补充核心谓词的状语块、补语块; 此外, 一些修饰或补充性成分并不与核心谓词比邻, 而是游离在主宾语前后, 但这种成分除了少数只出现在句首的状语外, 都可看作修饰、补充说明核心谓词的成分, 本文将其称为 “句饰语”, 也是状语、补语的一部分. 需要说明的是, 本文认为述语是句子核心, 不可或缺, 除了独词句外, 每个句子都有述语, 缺省述语的情况下将补出述语空位; 而对于多谓词结构连用, 如复谓句中的多谓词结构连用, 则认可该句有多个谓词核心; 对充当主语与宾语的谓词性结构, 进行进一步的递归分析. 这与其他类似树库有所区别, 如 CTB 强调每个句子都有主语, 而本文则强调每个句子都必须有述语; 而强调谓词性主语与宾语的递归分析也是与语块库的显著区别, 谓词性组块的递归分析, 避免了组块分析过于平铺、笼统, 有助于更准确地进行句法分析、后续扩展标注; 然而, 在本阶段虽然强调了述语的中心谓词与状语、补语边界, 但未对状语与补语内部作进一步分析, 容易导致状语与补语冗长, 使得一些重要句法语义信息不突出, 这是后期扩展阶段需要解决的问题.

主语块与宾语块是相对述语的句法位置的, 因其分布在述语两侧, 称为 “主宾语”, 其内部差异由其与述语块的相对位置标识.

衔接性成分和辅助性成分与述语在结构上, 并不直接相关. 前者起着连接上下文的作用, 后者帮助表达语气与状态、引起话题注意等, 分别认定为 “衔接语”、“辅助语”.

衔接语除了传统单复句理论中篇章连接词外[16] (这部分词较封闭, 大约在 450 个左右), 也包括一些难以进行句法分析但起显明连接作用的标记成分, 系统功能语法中以元话语标记来分析具有篇章组织功能的成分, 这是一种结合紧密、高频使用、主要表达程序意义、具有语篇组织功能及人际功能的语言单位. 徐赳赳[17] 将书面形式的元话语概括为词汇元话语、标点元话语和视觉元话语. 词汇元话语主要由词汇化、语法化的话语标记充当, 话语标记的语篇功能, 有着与传统连接词相似的语篇连接作用、句法功能边缘化的特点[18–19], 因此本文也将其看作一类衔接组块, 这类词很难精确个数, 但往往以某类构式为标记, 如 “代词 + 感官类动词”, 汉语中常用的话语标记大约在 400 个左右. 此外, 标点元话语起连接作用的主要是一些插补成分. 在篇章行文过程中, 临时插入到词串序列中, 以补充、解释、说明、强调前面提到的内容, 但并不与前后语言单位有句法结构关系, 本文也将其看作一种衔接组块, 以 “尤其是” “特别是” “如”最为常见, 在书面上, 这些插补成分往往前后有逗号、破折号, 或者被括号 “( )、[ ]”封闭标识出来. 事实上, 一些树库, 如 CTB、TCT、PCT, 在构建过程中, 已有意识地将部分高频出现的元话语在性质与功能上加以区分, 但并未对这类成分进行系统探讨, 通常优先将其分析为状语、谓语等, 只有实在难以进行句法分析的成分, 才单独分析, 主观性较大.

辅助语主要包括语气词、语气辅助语、呼应词、感叹词、拟声语、“是……的”等; 其中, 语气辅助语是指原本具有实在意义的词或词组在句中语义泛化, 不带后续成分, 不具言说义, 成为非语义重心, 发生词汇化、语法化, 帮助表达语气的词或词组 (如: “我们去食堂好不好/行不行/对不对?”); “是……的”中这种 “是” “的”, 作用是指明全句焦点或强调谓语、表达对主语的评议、叙述、描写. 其归属问题, 学界莫衷一是, 本树库将其处理为辅助语, 目的并非为其最终定性, 而是为了标引出来, 以便后期根据需要进行合适的处理; 其他主要是指用在谓词后, 不做述语的助词, 如 “罢了”、“而已”等.

人工标注时, 只需要根据各组块的定义, 以几个简单的标注符号直接标注. 述语组块方面, 以 “()” 标注述语及核心谓词的左右边界, “{}”标注主谓谓语句的谓语的左右边界, “[]”标注句饰语; 主宾组块方面, 以 “||”标注双宾语之间的边界, “{}”也标注谓词性主语与宾语, 而体词性的主宾语因其他组块边界的区分而自然得以区分; 以 “<>”标注衔接组

块; 以 “<<>>”标注辅助组块. 如 “<但是>多数人 (还是 (受) 不了) 这份苦<<啊>>. ”, 通过组块的界定, 以及分类的标注符号, 专业的标注人员能以更符合语言直觉的方式, 区分出衔接组块、主语块、状语块、核心谓词块、补语块、辅助组块.

此外, 以 “|”作为增加句边界的符号, 以 “\\”作为取消句边界的符号 (详见第 2.1.2 节); 对文句错误, 难以分析的句子, 以 “** **”标注错句左右边界, 内部不再进行分析.

至此, 本文便以符合语言直觉的方式, 在可分析的线性序列上, 穷尽所有句子成分, 完整地刻画出了句子以述语中的核心谓词为中心的句子骨架. 人工标注的句子骨架可以通过代码无歧义地转化为组块状句法结构树, 也可部分地转化为组块状依存结构, 为后续树库扩展标注打下基础.

1.1.2 篇章中 “句”的层次与边界标记

篇章中的句法分析、结构分析, “句”都是最基本的分析单元. 然而汉语并无被广泛认可的 “句”定义, 篇章分析理论中, 普遍以句号、问号、叹号作为句边界, 将篇章切割为一个个待分析的句子, 但大规模非标注书面语文本中, 几乎所有断句标点都可能存在歧义性, 也存在缺失或缺省的情况; 而从句调以及完整性等语音、句法角度笼统地对 “句”定义, 又缺少理论依据, 进而缺乏应用的指导意义[20]. 大规模树库构建, 如无系统的断句标准, 一方面影响标注者对各层级标注单元判断的一致性, 从而影响数据质量, 另一方面也不利于后期树库的使用与系统扩展.

从符合汉语基本原理及可计算的角度, 本文基本保持篇章原有段落结构, 在段落层面采纳了传统单复句的篇章理论, 将篇章中的 “句”首先分为单句、复句. 单句, 由句法成分、辅助成分、衔接成分构成, 独立表达完整语义, 与其他 “句”没有依存关系. 复句, 复句是由分句构成的表达一套完整语义的句子; 分句又由相对独立又互相依存的 “小句”和 “片段”构成; 小句由句法成分或辅助成分、衔接成分构成, 而片段是指能独立表达一定逻辑关系、施为义、情感态度义, 但需要语境才能较为独立表达意义, 一般由衔接成分、辅助成分充当. 因此, 自底向上地: 小句作为基本句单位, 可独立充当单句, 也可与其他小句形成分句关系, 共同构成复句; 一些片段主要起连接作用、语气作用, 也可看作一种特殊的分句; 而若干单句或复句构成段落, 若干段落构成篇章 (见图 2).

如此, 在短语结构的句法树上, 复句和单句都有根节点 (ROOT), 单句与分句都必有一层小句节点 (Independent phrase, IP), 分句可能还有一层片段节点 (Holophrastic phrase, HLP) (见图 2). 树库中有句层次标注的树库不多, CTB 从性质和功能角度, 描写了简单子句 (IP) 与复杂子句 (CP)、祈使语气 (IMP) 与疑问语气 (Q); TCT 与 PCT, 则以整句 (ZJ)、复句 (FJ)、单句 (DJ)、句群 (JQ) 描述句子间组合层次; 此外, TCT 也对句间关系进行了标注; 本文句的描写与 TCT、PCT 更接近, 借鉴了传统单复句理论, 对句间组合层次作了必要的标识, 对句子类型、句间关系的标识目前并未涉猎, 将留待句间结构标注完成, 不过对各层级句间边界的判断、衔接成分 (其他树库称为语篇成分、篇章连接成分等) 的判断标准更为明确客观, 从而降低了主观随意性.

有了对篇章中 “句”的层次结构区分后, 就可对各层次的句设定更为清晰的边界标准了. 本文综合前人研究成果, 归纳了 3 条允许标注者增加或删除句边界的标准: “终止性停顿标志” “互不作句法成

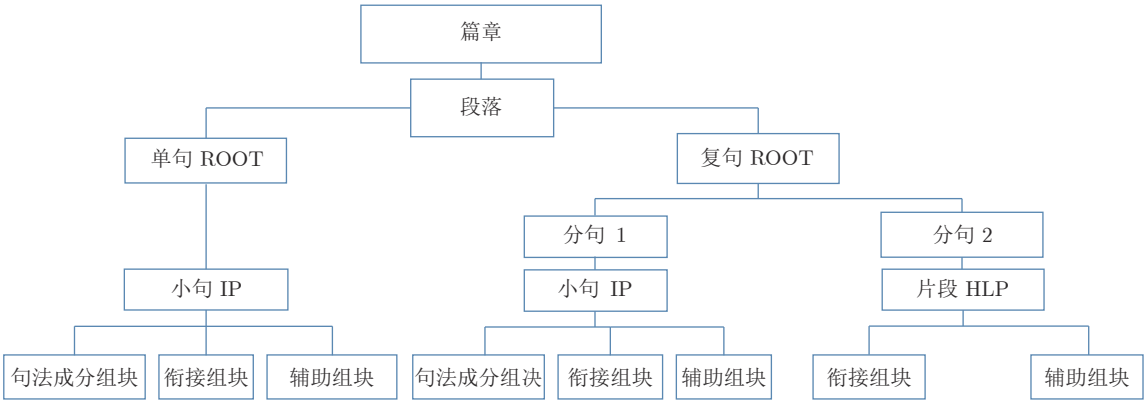


图 2 篇章中 “句”的层次结构示例

Fig. 2 A Sample of the hierarchy in Chinese discourse sentences

分” “表述功能” (限于篇幅, 仅简要介绍, 具体论证将另行他文).

终止性停顿标志, 主要指有断句作用的标点符号及一个标点句中多个谓词所支配的主语不同的情况.

标点符号往往是篇章中句边界的显性标记, 根据上述 “句”层次, 本文将句号、问号、叹号、省略号、分号、冒号作为可标记单句、复句边界的标点符号; 逗号、破折号作为标记分句边界的标点符号; 这 8 个标点符号统称为标句点号, 是一种重要的终止性停顿标记. 在分句时, 首先默认标句点号切句功能, 在此处断句, 但遇到并非表达成句作用时再根据其他标准取消其断句功能.

在一个标句点号所分割出来的句子内, 可能包含多个述语组块 (形式上如同复谓句), 而述语组块所支配的主语块并不相同, 即在连续的无标句点号的词串序列中, 包含、隐含着两个不同的独立主语或主语只在后一个述语块前出现, 形成两个较为独立的命题, 如例句 1 ~ 3 中, 竖线比邻的词组分属两个不同的命题结构, 包含或隐含着两个迥异的主语. 如果不加以分割, 无论是句法分析还是后续话题结构标注都会受到影响. 这种包含两个命题结构的词串序列, 前一组合结构后的宾语或后一组合的主语, 本文称其为 “换主语”标记, 是一种终止性停顿标记. 这种情况往往出现在 “V_1NV_2”词串序列中, 以及非规范, 或错误使用标点造成标句点号缺省的词串序列中.

例句 1. 刚想拿起来读 | 书就掉了.

例句 2. 内线 (打不开) | (赢不了) 对方!

例句 3. (不 (发酵))| 面包 (会比较 (干)).

标句点号与 “换主语”标记作为两种 “终止性停顿标志”作用互补. 标句点号针对大概率的句边界显性标记, 而 “换主语”标记是针对标句点号缺失的情况设定的句边界.

标句点号的不可靠性不仅表现为可作句边界的标点缺位, 也往往表现为语音停顿与切句停顿不一致的矛盾. 此时, 需要从句法上加以区分: 线性序列上, 标句点号比邻的两部分内容有主谓结构、状中/定中结构、述补结构关系, 则认为标句点号比邻的两部分是一个单句或一个分句, 否则是两个单句或两个分句. 这种判别标准称为 “互不作句法成分”, 用以解决标句点号歧义问题, 如:

例句 4. 5G (是) 在 4G 基础上\\, 把人的连接拓展为物的连接.

例句 5. 13 亿多中国人{, 一个 (都不能 (少))}!

“互不作句法成分”在面对体词性、谓词性组合与标句点号比邻的句子有主谓关系时, 依然会有难以判定的情况, 如图 1 中的例句增加逗号 “后来, 机场建设征用农田, 引起农民强烈反对, ……”如果仍看作一个小句, 那么假如第二个逗号后还有多个类似结构与之并列, 这个 “小句”可能会过长, 结构嵌套过深, 不利于人标注, 也影响句法分析效果; 而在理论上也有学者指出, 谓词性组合具有表述性, 而体词性组合表述性不强, 一般不充当分句[21]; 由于动转名可以无标记的方式进行, subject–verb–object (SVO)、subject–verb (SV)、verb–object (VO) 结构可作话题主语, 然而这种句式话题与述题关系松散, 可以重新分析为复句[22]. 因此本文据此设定了 “表述功能”的切句标准: 比邻的两个词组, 前面的谓词性组合与后一组合有主谓关系时, 将其看作两个小句 (例句 6); 标句点号比邻的组合序列是体词性与谓词性组合, 且二者构成主谓关系, 则看作一个小句 (例句 7); 标句点号比邻的组合序列都是体词性组合, 且二者构成主谓、并列关系, 则将其看作一个省略了述语的主谓句 (例句 8); 如没有主谓、并列关系, 此类句式有明显的描绘性语义, 主要用于表达形象感, 则认为各自独立自足 (例句 9).

例句 6. {考试 (考砸了)}(让) 爸爸 (很 (失望)).
考试 (考砸了), (让) 爸爸 (很 (失望)).

例句 7. 五年的时光\\, (让) 他 (学会了){(游泳)、(捞) 鱼、(养) 鸡、(喂) 鸭}.

例句 8. 粒径\\, () 泥沙颗粒大小的一种量度.

例句 9. 蓝蓝的天空, 洁白的云朵, 高高的白杨树, 明亮耀眼的阳光, 一望无际的草场, 这 (就 (是)) 我憧憬的蒙古科尔沁大草原<<呵>>.

1.1.3 树库标记集

根据第 1.1.1 ~ 1.1.2 节所述的标注体系, 本文的树库标记主要包含组块性质标记、组块功能与用途标记、句边界标记. 表 1 中, 第 1 ~ 4 为组块性质标记, NP 与 VP、UNK 主要用于描述主宾语、述语整体及核心谓词的体谓性, 其他成分的组块性质不作分析, 以 NULL 标记; 第 4 ~ 11 为组块的功能与用途标记, 功能是从句法上描述组块单元充当句法角色的情况, 用途是从语篇功能与人际功能描述组块单元的连接作用、语气作用; 第 12 ~ 15 为句边界与层次标记.

1.2 语料选择

本树库以大规模多领域篇章语料为标注对象, 前期以人工标注为主, 分阶段、分语料层次逐步加入机器标注的方式构建完成. 根据用词与句法结构复杂度, 标注语料类型由简单到复杂逐步过渡, 以

表 1 树库标记集
Table 1 Tags for chunk-tree

序号	符号	标记类型
1	VP	谓词性组块
2	NP	体词性组块
3	UNK	谓词与体词并列的组块
4	NULL	其它性质组块
5	PRD	述语
6	NPRE	名词谓语
7	MOD	状语、补语
8	SBJ	主语
9	OBJ	宾语
10	CON	衔接组块
11	AUX	辅助组块
12	ROOT	单复句
13	IP	完整小句
14	HLP	独词句或片段
15	W	标点

便打磨规范、开发标注平台、训练标注人员、开发自动标注工具. 而前期以百度百科、新浪和新华社新闻、国家专利为标注对象: 百科与新闻涉及到社会生活的各方面, 内容丰富、结构严谨、用词造句的风格多样, 但并不复杂且较为标准, 可以作为树库的基础数据, 而专利文本是典型的技术说明型、法律型文本, 用词偏晦涩、结构虽然严谨但复杂, 涉及技术领域广, 可以覆盖较多技术性文本的特点; 后期用已标注的数据训练机器标注模型, 采用机器标注, 人机校对的模式, 标注其他语体文本, 并不断用已标注的语料训练迭代机器标注模型, 以标注模型表现的领域稳定性决定某一领域的语料规模; 此外, 注重长尾句型, 根据机标错误句型分布, 动态增减相应句型的语料.

2 标注工程实施

2.1 研制标注规范

制定标注规范是树库构建过程中最重要也是最困难的任务. 科学、完善的标注规范, 对数据符合预期目标、保证标注的一致性和准确性至关重要. 在规范起草与修订过程中, 为保持规范与预期目标一致, 本树库有三条主要原则: 1) 以述语为核心, 无歧义呈现句子骨架; 2) 人工标注及句法树合格性校验时, 不做分词及词性处理, 切分后形成组块状的组合序列, 训练机器标注模型时, 则在组块分析的基础上, 块内分词及词性标注; 3) 平衡规范的稳定性与动态修订性, 规则分级, 上级稳定, 下级慎重动态增减, 例句动态详举.

汉语中存在身兼数职、功能复杂的实词或虚词; 同时, 进入篇章中的句子, 语序灵活, 意合现象普遍, 语用层面的 “经济性”原则被普遍使用, 反过来促使语序、省略、类比、并列等变得灵活, 往往导致在篇章层面标注的句法结构与语义结构错配, 句法分析得不到语义、语用的支持. 为此, 本树库坚持 “以述语为核心, 无歧义呈现句子骨架”的原则, 即, 参考上下文意, 使标注的句子骨架能尽可能无歧义地表示组块间依存关系, 反映实际支配关系, 也为后续成分缺省与共享标注任务准备高质的带标数据. 如:

例句 10. 我<<是>>(历来 (主张)){军队 (要 (艰苦奋斗))}<<的>>.

这书 (是) 她的.

例句 11. 这孩子 [跟狐狸一样,] (很 (狡猾))

今年的题 (跟去年 (一样)).

例句 12. 电磁流量计密封性能 (好), (还可 (用于)) 自来水和地下水道系统. <而且> (测量过程不与流体 (接触)), (适于) 制药、生物化学和食品工业.

例句 13(a). (给) 你 (拿) 件毛衣 [来](如何)?

例句 13(b). [不用](), [不用](), 我 (不 (冷)).

例句 10 中起焦点标记作用的 “是”与系动词的 “是”区分开来, 一个以述语符号标注, 一个以辅助组块符号标注; 例句 11, 都是 “跟……一样”, 一个表示比拟, 一个表示比较, 标注出来的句子骨架也有所区别; 例句 12, “测量过程”, 在单句层面是主语, 但从上下文看, 并非 “测量过程”不与流体接触, 而是在 “测量过程”中, “电磁流量计”不与流体接触; 例句 13(b) 实际省略核心谓词 “拿”, 这在对话中是常见情况, 即, 对话题进行删除或省略, 主要强调连句成篇、约束焦点、触发预设、表达立场或情感和态度、人际互动等作用, 而将述语缺省, 因此在本树库规范中规定需依据实际情况补出述语空位.

“不做分词及词性处理, 切分后形成组块状的组合序列”, 主要针对通用性、固定表达, 不做进一步切分. 熟语、固定表达、古语引用、古语用法、公式结构等, 内部结构高度凝固成为一种习用的、结构相对固定的叙述性语言单元, 对这部分表达尽量大颗粒度标注.

最后, 标注规范是对树库设计思想的具体体现, 语言现象的复杂性, 决定了任何一种标注规范都无法一步到位, 需要在标注实践过程中不断充实、修订、完善, 但无规划的修订, 会导致整个标注工作缺乏系统性, 增加标注与标注平台开发成本. 本项目在经过长达 5 个月的规范制定工作后, 起草了初版

标注规范, 并确定了 “规则分级, 上级稳定, 下级慎重动态增减, 例句动态详举”的增订原则, 将规则按类逐层细分, 上级规则稳定不易变动, 根据实践以增加详细说明的细则, 而对每类细则详举典型或需复杂判断的标注例句, 以减少标注者主观判断, 提高标注一致性, 目前规范中有 11 章、87 节内容、900 多条例句, 除词表附录共 66 页. 通过这样的策略, 可以尽力平衡规范前期的稳定性与动态修订性, 同时这样的编排也便于标注者索引.

2.2　标注人员与标注质量、速度控制

平衡标注质量、速度与规模, 需要高专业素养的标注人员、有效的质量评测与反馈机制、高效的标注与管理办法.

句法结构标注的难度远高于其他标注任务, 对标注人员的专业素养要求高. 本文始终维持着一个语言学专业的标注团队, 参与标注的团队成员都需通过标注考核, 每期有效标注人员保持在 35 人以上.

在质量评测与反馈方面, 本文主要做了以下工作:

1) 改进计算标注一致性的 Kappa 算法, 以精确衡量每个组块及全文本标注质量, 为组织标注工作、数据使用提供参考. 句法结构标注的一致性, 本质是递归结构切分与定性的一致, 结合自身标签设计特点以及参考 Holle 等[23] 相关工作, 将比对文本中每个字符依次从 0 开始编码, 每个非标注字符都有一个唯一的、有序的起止位置编码, 每两两标注一致的字符起止位置对齐, 嵌套结构按起止码由大到小的顺序排列, 中间按组合顺序依次排列, 一方缺失的补空, 标注符号与起止位置都相同即为一致的判定标准. 据此按标注符号构建列联表, 计算全文本单个标注符号 Kappa 系数, 再根据各标注符号在文本中的占比为权重, 计算全文本 Kappa值, 计算公式为 (具体算法推导超出本文范围, 将另行他文详述):

$$K=\sum_{i=1}^{10}\rho_i k_i$$

如第 3.1.1 节所述, 本树库在人工标注阶段, 有 10 类人工标注符号, i 为遍历时的第 i 类标注符; k_i ($-1\leq k_i\leq 1$) 为每类标注符 Kappa 值. ρ_i 为各类标注符在全文本标注符中所占比重, $\rho=\sum_{i=1}^{10}\cdot\rho_i=1$. 因计算目的不同, 权重 ρ_i 有两种计算方法. 下面用 m_i 表示当前统计标注符号在两个人标注的文本中, 总共出现的次数, 用 n 表示 2 个人各自标注的符号数之和, 即, $n=\sum_{i=1}^{10}m_i$. 两种计算方式如下:

$$\rho_i=\frac{m_i}{n} \tag{1}$$

$$\rho_i=\frac{\lg(1+m_i)}{\sum\limits_{i=1}^{10}\lg(1+m_i)} \tag{2}$$

式中, 对数加 1 以防止出现 lg0 的情况. 权重 (1) 为计算文本质量的方法, 权重 (2) 用以作为评估标注人员标注表现进而计算工资的一种方法, 从而鼓励标注者重视低频标签的标注准确性.

2) 按 20% ~ 30% 的比例, 人员与语料由系统双随机派发 “埋雷文本”与 “审核文本”. 每一篇 “埋雷文本”需要两位标注人员在得到第 1 次一致性校验结果后, 讨论协商后再次提交进行一致性校验, 难以一致的问题提交管理人员仲裁; 埋雷文本第 1 次所得 Kappa 值, 作为该标注人员工资系数标准, 也作为该标注人员当期所标注的其他 “审核文本”是否进入树库备选的依据 (需 Kappa ≥ 0.8); “审核文本”则是不进行一致性校验, 但由管理人员随机抽查的文本, 通过审核的 “审核文本”的字数总和, 决定标注人员的工资总额, 而不合格的文本直接全文本报废处理, 管理人员同时也通过程序抽取出的句子骨架, 重点审核低频出现的句子骨架, 加以反馈修订. 所有备选树库中的数据需要再经过树结构转换代码检验, 不合法的句法树将被抛出, 合格的句法树构成了最终的树库.

3) 开发通用型在线标注平台、管理计算工具 (见图 3). 项目组研发了一个标注管理平台, 并在实践中不断完善. 在浅层句法标注上, 符号标注比树图可视化编辑更直观便捷, 而有层次嵌套的结构, 符号标注比色块标注更便利, 但色块标注更便于标注检查; 因此, 平台支持选择性输入与快捷键输入, 支持以色块和标注前后比对进行检查; 此外, 通用标注平台也支持标注相关人员协调管理, 对标注行为进行实时跟踪、反馈, 为标注任务管理提供参考; 集成的管理工具可辅助、补充在线管理系统, 保证管理灵活性的同时, 确保管理的系统性.

2.3　人机互助的语料加工

在完成一定规模数据标注后, 初步分类训练了机器标注模型, 初步探索了 “机标人校”的标注模式 (具体实验另行他文介绍). 本文以已标注的百科、新闻及专利语料为训练语料, 采用自注意力机制编码[24] 和基于 Cocke-Younger-Kasami (CYK) 的图表算法解码[25], 以 Bert[26] 进行预训练, 训练所得句法分析器的 F1 值为 94.3; 其次, 以句法分析器自动标注新闻、百科、专利以及小学生作文、法律判决书、

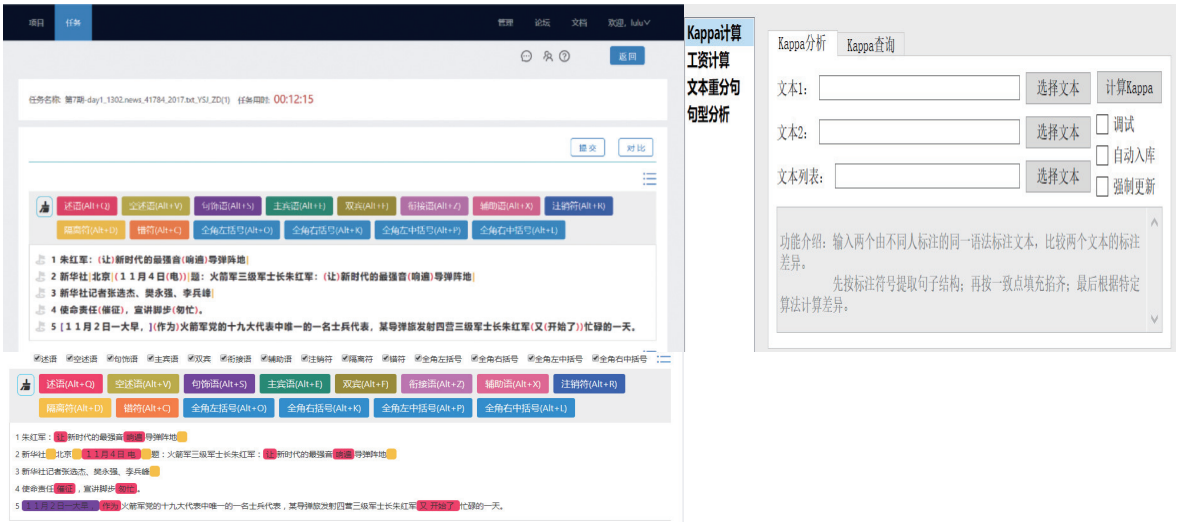


图 3 标注平台标注界面及管理工具界面

Fig. 3 The interface of annotation website and management tool

科技说明文的测试文本; 最后, 将机器标注的结果与人工标注的结果进行一致性校验, 经过多次校订的人工标注与前三类文本校验 Kappa 值分别为 0.835、0.834、0.639, 而后三类文本与人工标注一致性校验的 Kappa 值, 分别为 0.70、0.63、0.83. 本文根据机器标注的表现, 对于机器标注已能满足使用需求的文本领域 (使用需求根据任务定义), 则不再进行人工标注; 对机器标注还不能满足使用需求的领域, 则采用机器标注、人工校订的方式增加该领域训练数据, 不断迭代标注模型的领域迁移能力, 以最大限度减少了人工投入, 丰富树库语料领域.

3 树库统计分析

3.1 标注数据分布

截至到目前为止, 以根节点计, 树库中已有 27.8 万句, 见表 2, 其中 47% 为新闻, 28% 为百科, 25% 为专利 (扩容中). 新闻数据涉及领域广泛, 因此语料占比较高, 专利数据将陆续加入新数据. 所有数据中, 两类黄金数据占总数据的 25.5%, 随着树库扩容, 黄金数据会增加, 但比例变化较小. 根节点 (ROOT) 字长分布也在一定程度反映了文本难度, 统计数据也验证了这一点, 三类文本平均字长由小到大依次为百科、新闻、专利文本.

以小句 (IP) 为对象, 从 67.4 万余条 IP 中仅抽取出 3 万余条句型骨架, 其中有 9 408 条句型骨架包含至少两条例句, 而 10 余条句型骨架却涵盖树库中绝大多数小句 (图 4 展示了 9 408 条实例句大于 1 的句型框架频次与排名顺序双自然对数分布), 低频句型骨架涵盖小句数量少, 自身条数却较大. 后期树库构建需要据此有针对性地扩充低频句型相关语料; 与此同时, 根据这些句型骨架, 已初步构建

表 2 目前有效标注语料分布

Table 2 The data distribution of the valid annotated data

类别	ROOT	IP	HLP	汉字数	文件数	说明
新闻	132 009	359 373	50 378	4 920 170	4 813	新浪 2006、新华社新闻 2012 ~ 2018 年新闻
百科	76 595	149 097	14 823	2 376 151	2 982	自动化控制系统、电子学与计算机、轻工、大气与海洋及水文科学、航空航天、经济学
专利	69 260	166 462	16 966	2 839 935	3 915	2018 年国家专利申请文书描述与权利申明部分
	ROOT 中汉字字长分布					IP 中字长分布
类别	平均	最大	最小	中位数	众数	平均
新闻	37	837	1	33	1	13.69
百科	31	251	0	27	20	15.94
专利	40	819	0	29	16	17.06

一个提供结构检索、词、词性混合检索的句型库, 以供语言学研究使用, 如以检索式 “%(.v 以){@}”检索 “‘动词/动语素’+‘以’”后跟谓词宾语的所有结构, 目前库中可检索到 157 条记录, 比如: “被告人吕宏(予以){应允}. ” “权钱交易、权力寻租现象 (得以){防控}. ” 等, 而随着自动标注模型的标注准确性提升, 可以自动分析任何语料, 以提供更多结构检索实例, 为语言学研究服务.

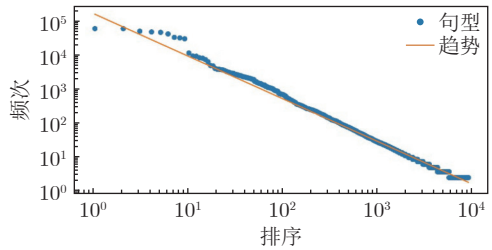


图 4　基本句型随机法齐夫对数分布

Fig. 4　The rank-frequency random logarithmic distribution for the sentence patterns

此外, 从小句标点使用情况的分布来看, 破折号、叹号、句号在较为正式的文本中做句边界的置信度非常高 (图 5), 分号、省略号次之, 冒号、逗号、问号作句边界, 则需要谨慎, 其中问号的情况较为复杂, 有不识别的符号转码造成, 也有本身使用有误的情况; 然而在句边界识别问题上, 最突出的问题不是标句点号的歧义性, 而是缺省标句点号, 以空格代替是常见情况, 少部分没有任何显性句边界标记, 需以 “换主语”标记进行判断.

3.2　树库标注质量评估

一致性校验是衡量树库质量和标注难度的可靠指标, 也是衔接本树库构建过程中各环节的最重要的数据参考. 由于本树库构建以全文本篇章为语料, 因此在决定文本是否进入备选树库时, 以全文本

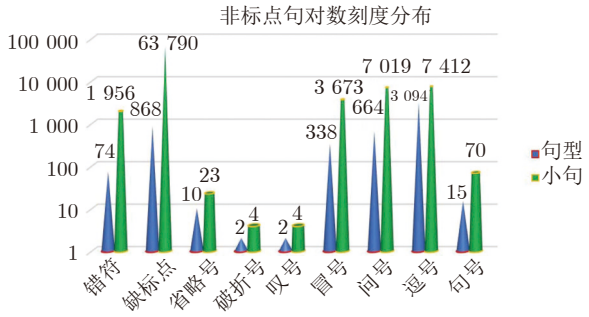


图 5　标句点号失效、缺省在 IP 小句中的分布

Fig. 5　Sentence-division-punctuation is invalid or missing

Kappa ≥ 0.8 为标准 (见图 6). 目前, 平均 Kappa 值为 0.87, 各主要标注符号的一致性校验也均在 0.8 以上, 其中 13% 的文件 Kappa 值超过 0.95, 约 24 万余字规模; 除了文件总体的一致性校验合格, 大部分成分块及边界判断的一致性校验也是合格的, 而句边界 (注销符) 及谓词性主宾语、主谓谓语一致性校验结果较低. 一方面, 这些边界及成分块判定往往是标注的难点, 其内部较为复杂, 主观判断较多, 另一方面也是后期本文要着重增加的长尾句型.

4　结束语

本文介绍了一种基于篇章的、便于多层次结构扩展标注的浅层句法标注体系, 并据此, 初步构建了一个千万汉字级的浅层短语结构树库, 提出一种以述语为核心的句子骨架标注体系, 有助于保证质量的情况下, 进行大规模、多层次结构标注. 同时也探索了众包环境下高效标注管理模式, 为后续各项扩展任务奠定了基础; 未来本文将从三个方面对树库进行扩容: 1) 依据解析器分析句型骨架效果、目前树库句型骨架分布抽取相应的篇章、段落、句子进行有针对性的标注, 丰富树库中低频句型语料,

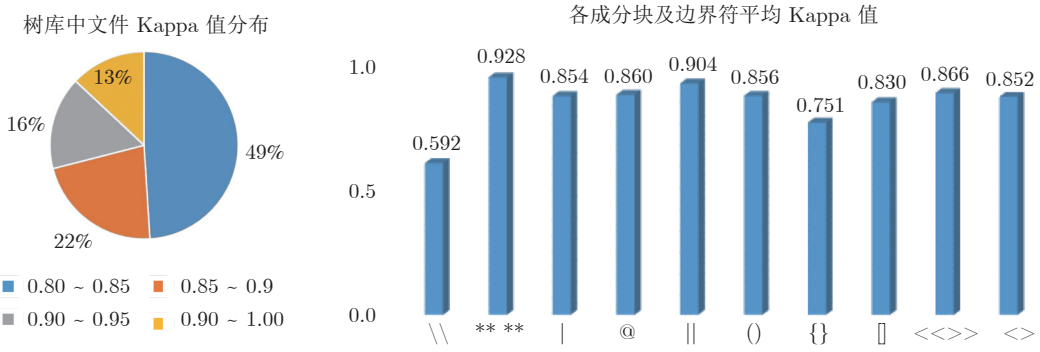


图 6　树库全文本 Kappa 值分布与各标注符号 Kappa 平均值

Fig. 6　The distribution of Kappa coefficient of the text in the Treebank, and the mean of every kind of label′s Kappa

同时依据模型标注准确率的稳定性决定语料类型在树库中的比例; 2) 对已标注 Kappa 值较低的文本进行人机互助的二次复标、审校, 完善人机协同标注模式、全面开启人机协同标注, 加快树库构建速度, 提升预标注模型性能; 3) 对已构建的句法块树库, 进行组块依存与话题结构标注, 开发依存树库解析模型, 进一步完善本树库, 从而构建完整的篇章块依存树库, 为后续延展任务打下基础.

References

1 Zhang X, Xue N. Extending and scaling up the Chinese treebank annotation. In: Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing. Tianjin, China: 2012: 27–34.

2 Zhou Qiang, Zhang Wei, Yu Shi-Wen. The building of Chinese treebank. *Journal of Chinese Information Processing*, 1997, **11**(4): 43–52
(周强, 张伟, 俞士汶. 汉语树库的构建. 中文信息学报, 1997, **11**(4): 43–52)

3 Zhou Qiang. Annotation scheme for Chinese treebank. *Journal of Chinese Information Processing*, 2004, **18**(4): 2–9
(周强. 汉语句法树库标注体系. 中文信息学报, 2004, **18**(4): 2–9)

4 Chen K J, Luo C C, Chang M C, Chen F Y, Chen C J, Huang C R. Sinica Treebank: Design criteria, representational issues and implementation, Chapter 13 [Online], available: https://link.springer.com/chapter/10.1007%2F978-94-010-0201-1_13, April 16, 2019.

5 Che W, Li Z, Liu T. Chinese dependency treebank1.0 (LDC-2012T05) [DB/OL].Philadelphia: Linguistic Data Consortium [Online], available: https://catalog.ldc.upenn.edu/LDC2012 T05, April 16, 2019.

6 Guo Li-Juan, Pen Xue, Li Zheng-Hua, Zhang Min. Construction of Chinese dependency syntax treebanks for multi-domain and multi-source texts. *Journal of Chinese Information Processing*, 2019, **33**(2): 34–42
(郭丽娟, 彭雪, 李正华, 张民. 面向多领域多来源文本的汉语依存句法树库构建. 中文信息学报, 2019, **33**(2): 34–42)

7 Brody M. Phrase structure and dependence [Online], available: http://real-eod.mtak.hu/8176/1/WorkingPapersInTheTheoryOfGrammar_01-1_1994.pdf, April 16, 2019.

8 Qiu Li-Kun, Jin Peng, Wang Hou-Feng. A multi-view Chinese treebank based on dependency grammar. *Journal of Chinese Information Processing*, 2015, **29**(3): 9–15
(邱立坤, 金澎, 王厚峰. 基于依存语法构建多视图汉语树库. 中文信息学报, 2015, **29**(3): 9–15)

9 Zhou Qiang. Build a large scale Chinese functional chunkbank. In: Proceedings of the 6th National Conference on Computational Linguistics-natural Language Understanding and Machine Translation. Taiyuan, China: 2001. 6
(周强. 构建大规模的汉语语块库. 自然语言理解与机器翻译—全国第六届计算语言学联合学术会议. 太原, 中国: 2001. 6)

10 Xue N, Palmer M. Adding semantic roles to the Chinese treebank. *Natural Language Engineering*, 2009, **15**(1): 143–172

11 Kong F, Wang H L, Zhou G D. Suvery on Chinese discourse understanding. *Journal of Software*, 2019, **30**(7): 2052–2072

12 Qian Xiao-Fei. Research review on chunk parsing. *Modern Chinese*, 2018, **2018**(6): 166–170
(钱小飞. 组块分析研究综述. 现代语文, 2018, **2018**(6): 166–170)

13 Chu C, Nakazawa T, Kawahara D, et al. SCTB: A Chinese treebank in scientific domain. In: Proceedings of the 12th Workshop on Asian Language Resources. Osaka, Japan: 2016. 59–67

14 Zhao Chun-Li, Shi Ding-Xu. Mood, modality and sentence type. *Foreign Language Teaching and Research*, 2011, **43**(4): 483–500, 639
(赵春利, 石定栩. 语气、情态与句子功能类型. 外语教学与研究, 2011, **43**(4): 483–500, 639)

15 Hu Zhuang-Lin. *Discourse Cohesion and Coherence*. Shanghai: Shanghai Foreign Language Education Press, 1994: 108–109
(胡壮麟. 语篇的衔接与连贯. 上海: 上海外语教育出版社, 1994: 108–109)

16 Xing Fu-Yi. *The Research on Chinese Sentences With Two or Mo-re Clause*. Beijing: The Commercial Press, 2001: 2–6, 26–31, 38–56, 546–548
(邢福义. 汉语复句研究. 北京: 商务印书馆, 2001: 2–6, 26–31, 38–56, 546–548)

17 Xu Jiu-Jiu. *The Text Linguistics of Modern Chinese*. Beijing: The Commercial Press, 2010: 218–222
(徐赳赳. 现代汉语篇章语言学. 北京: 商务印书馆, 2010: 218–222)

18 Li Xiu-Ming. The Research of Chinese Metadiscourse Marker [Ph.D. dissertation], Fudan University, China, 2006
(李秀明. 汉语元话语标记研究. 复旦大学, 中国, 2006)

19 Yang Yi-Fei. Connection in Modern Chinese Discourses [Ph.D. dissertation], Fudan University, China, 2011
(杨一飞. 语篇中的连接手段. 复旦大学, 中国, 2011)

20 Song Rou, Ge Shi-Li, Shang Ying, Lu Da-Wei. Chinese sentence and clause for text information processing. *Journal of Chinese Information Processing*, 2017, **31**(2): 18–24, 35
(宋柔, 葛诗利, 尚英, 卢达威. 面向文本信息处理的汉语句子和小句. 中文信息学报, 2017, **31**(2): 18–24, 35)

21 Chen Rong-Chun. Talk about Chinese compound sentence from declarability. *Linguistic Researches*, 1981, **1981**(1): 46–51
(陈荣春. 从句子的表述性谈单句复句的划分. 语文研究, 1981, **1981**(1): 46–51)

22 Dong Xiu-Fang. *The Phenomenon and Regularity of Chinese Lexicalization and Grammaticalization*. Shanghai: Akademia Press, 2017: 143–144
(董秀芳. 汉语词汇化和语法化的现象与规律. 上海: 学林出版社, 2017: 143–144)

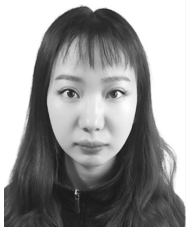
23 Holle H, Robert R. *Understanding Body Movement: A Guide to Empirical Research on Nonverbal Behaviour With an Introduction to the NEUROGES Coding System*. Frankfurt am Main: Peter Lang GmbH Internationaler Verlag der Wissenschaffen, 2013. 261–277

24 Kitaev N, Cao S, Klein D. Multi-lingual constituency parsing with self-attention and pretraining [Online], available: https://arxiv.org/abs/1812.11760, April 16, 2019.

25 Mitchell S, Jacob A, Dan K. A minimal spanbased neural constituency parser. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada: 2017, 818–827

26 Devlin J, Chang M W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, 2018, arXiv: 1810.04805

卢 露 北京语言大学信息科学学院硕士研究生. 主要研究方向为语言智能与技术.
E-mail: 201821198367@stu.blcu.edu.cn
(**LU Lu** Master student at the College of Information Science, Beijing Language and Culture University. Her main research interest is linguistic intelligence and technology.)



矫红岩 北京语言大学信息科学学院硕士研究生. 主要研究方向为自然语言处理.
E-mail: jiaohongyan0815@163.com
(**JIAO Hong-Yan** Master student at the College of Information Science, Beijing Language and Culture University. Her main research interest is natural language processing.)

李 梦 北京语言大学信息科学学院硕士研究生. 主要研究方向为计算机应用技术.
E-mail: limeng_gertrude@163.com
(**LI Meng** Master student at the College of Information Science, Beijing Language and Culture University. Her main research interest is computer applications technology.)

荀恩东 北京语言大学信息科学学院教授. 主要研究方向为自然语言处理. 本文通信作者.
E-mail: edxun@blcu.edu.cn
(**XUN En-Dong** Professor at the College of Information Science, Beijing Language and Culture University. His main research interest is natural language processing. Corresponding author of this paper.)