**文章编号：**

# 融合从底向上与自顶向下的中文复杂句人工标注方法*

**摘要：**大规模、高质量的中文树库的建立对中文句法分析的发展有着重要的意义，但是对于字数多、结构层次复杂的复杂句的标注仍费时费力，严重影响了树库的建立速度，以及中文句法分析的发展。本文提出了一种融合了从底向上和自顶向下的复杂句标注方法。首先，利用块切分辅助工具，进行句子的语法成分边界切分，对切分后的块做去重和筛选处理；其次，利用CCC Parser对块进行自动分析；再次，利用块校对辅助工具，对分析器处理过的块进行人工校对；然后，进行块还原，再利用块校对辅助工具，进行针对整句的人工校对；最后，由校审专家确定最终质量。实验表明，与从底向上的标注方法相比，本文方法的校对速度是传统方法的2倍多，且整体差异率和分阶段的差异率降低了约20%，证明本文方法在对中文复杂度的标注是有效的。

**关键词：**概念复合块；从底向上语料标注方法；从底向上和自顶向下语料标注方法

**中图分类号：**TP391　　**文献标识码：**A

# A Manual Annotation Approach to Chinese Complex Sentences by using Bottom-up and Top-dowm

**Abstract:** Large-scale development of the establishment of high-quality Chinese Treebank of Chinese syntactic analysis has important significance, but for more words and complex sentences complex hierarchy of labels is still time-consuming and seriously affecting the speed of the tree to build the library, as well as Chinese parsing development. This paper presents a blend of bottom-up and top-down complex sentences from tagging methods. First, using the block segmentation aids segment grammar component boundaries of the sentences, re-do and filter the blocks after segmentation; Secondly, CCC Parser analyzes the block automatically; once again, using the block proofing aids proofread manually the parser processed blocks; then, reduce the blocks and reusing the blocks proofing aids proofread the sentences manually; Finally, determine the final quality by the review expert. Experimental results show that, compared with the mark from the bottom up approach, proofreading speed of this method is two times more than traditional methods, and the overall difference in the rate and phased difference was reduced by about 20%, indicating that the proposed method for Chinese complex sentences the label is effective.

**Key words:** Concept Compound Chunk; bottom-up method of corpus annotation; bottom-up and top-down method of corpus annotation

---

* **收稿日期：**　　**定稿日期：**
**基金项目：**
**作者简介：**

## 1.引言

自然语言处理的分析技术可以分为两个层面，一个是浅层分析，如词法分析，词法分析一般只需对句子的局部范围进行分析处理[1]，目前已经基本成熟，其原因主要为针对词法分析有很多高质量、大规模的语料库的存在，使得基于这种涵盖绝大部分语言现象和特征的分析器可以达到比较理想的效果，其标志是它已经被成功地应用于文本检索、文本分类、信息抽取等方面[2]，并对这些应用产生了实质性的帮助。另一个是深层分析，如句法分析，句法分析是对一个汉语的句子进行深层次的、全局的分析与处理[3]。目前，国内外对此的研究还很不成熟，没有达到完全实用的程度。与词法分析相比，国内外针对句法分析的语料库很少，且标准参差不齐，规模不大，这直接制约了句法分析的发展。因此，进行高质量、大规模的语料标注变得异常重要。

英国的Lancaster-Leeds树库在1984年到1988年的五年间总共加工产生了二百多万词的树库语料，通过人工标注形成了一个4.5万词的小树库，通过一个概率分析器自动加工了14.4万词的语料，并进行了人工校对，从1987年起,他们开始尝试着采用生成较深（即具有较多中间结果）的分析树[4]。从1998年到2000年，宾州大学建成了宾州中文树库CTB-I，树库以新华社的10万词新闻文本为语料；2003年进一步完成了CTB-II的标注，树库增加了人民日报、香港新闻电讯和从其他语言翻译过来的中文稿件，规模为40万词[5]。清华汉语树库（TCT）是以从大规模的经过基本信息标注的汉语平衡语料库中提取出100万汉字规模的文本为语料，经过自动句法分析和人工校对，形成高质量的汉语句法树库语料[6]。

概念复合块（Concept Compound Chunk, CCC）是由2个或2个以上的词语按照一定的关联关系组合形成的信息描述单位[7]，它采用完全的二叉树结构。CCC标注的处理目标是对一句经过词语切分和词性标注处理的句子,分析出其中的不同实义词和功能词组合形成的概念复合块，确定其外部成分和内部关系标记，以此为基础，组合形成句子中描述不同事件内容的基本事件句式和变形事件句式描述序列,为进一步进行汉语句子的谓词论元关系分析打下基础[8]。

目前，语料标注的研究成为计算语言学领域的一个重要的研究方向。现在，很多人的研究集中在标注语料资源的开发以及建立一套标准的语料库标注体系[9]。语料标注是一个庞大的工程，它需要消耗大量的人力、物力和时间，且标注质量不一定能够达到人们的要求[10]。尤其是复杂句，较简单句来说，结构复杂，逻辑层次多，并列成分多。为了缩短语料标注时间和提高语料标注质量,本文融合了自顶向下和从底向上的语料标注方法对中文复杂句进行标注。

## 2.复杂句的特点

通过分析语料库中复杂句的主、状、谓、宾成分，以及复杂句的特殊句式，得到复杂句的特点如下：

（1）结构复杂，逻辑层次多

（2）常须根据上下文作词义的引申；

（3）常须根据上下文对代词的指代关系做出判断；

（4）并列成分多；

（5）修饰语多，特别是后置定语很长；

（6）习惯搭配和成语经常出现。

为了更好的描述复杂句的复杂程度，我们进行了如下定义：

**定义 1** 小句分析难度（Clause Processing Difficulty, CPD），为一个小句的人工标注处理以及机器自动分析的困难程度[11]。

针对小句的自动分析难度估计,主要考虑了汉语中可以找到明显形式标记的两类常用的小句内容复杂化操作：

并列复合结构：不同描述块通过顿号和并列连词(cC)组合形成的内容复合结构，以小句中出现的顿号和并列连词数量(CPN)作为分析难度定量估计依据；

子句嵌套结构：表示不同事件内容的子句嵌套形成的复杂层次结构，以小句出现的核心动词数量(KVN)作为难度估计依据。下面对其中的核心动词定义方法进行简要说明：

a）小句中的普通动词(v)作为核心动词；

b）名动词(vN)和助动词(wM)不作为核心动词；

c）趋向动词(v)（通过特殊趋向动词表判定），只有在小句中的动词数目为 1 时，才作为核心动词；在其他情况下，不作为核心动词；

这样，对每个小句，可以得到以下的小句分析难度(CPD)估计：CPD=CPN+KVN。

**定义 2** 句子分析难度（Sentence Processing Difficulty, SPD），为句子的人工标注处理以及机器自动分析的困难程度。

SPD 的处理方法为，对每个句子，首先排除其中所有词语长度为 1 的小句，得到其他有效小句（词语长度>=2）总数，然后计算其中全部有效小句的分析难度估计的算术平均值，作为该句子的自动分析难度估计值。

**定义 3** 平均分析难度（Averaged Processing Difficulty, APD），为全部句子的句子分析难度的算术平均值。

APD 的计算方法为，每个句子的句子分析难度总和除以句子总数。

相比于简单句，复杂句的复杂程度，如表 1、图 1 和图 2 所示：

**表 1 复杂句与简单句量化指标的比较**

| | 句数 | 总词数 | 平均词数 | APD |
|---|---|---|---|---|
| 复杂句 | 9087 | 497781 | 54.78 | 2.87 |
| 简单句 | 36348 | 781280 | 21.49 | 1.79 |

表 1 说明了复杂句和简单句的句数、总词数、平均词数和 APD。简单句的句数是复杂句的 4 倍，简单句总词数也比复杂句总次数多，复杂句的平均词数是简单句的 2 倍以上，平均句子分析难度也比简单句高很多。因此，可以看出复杂句比简单句复杂的多。

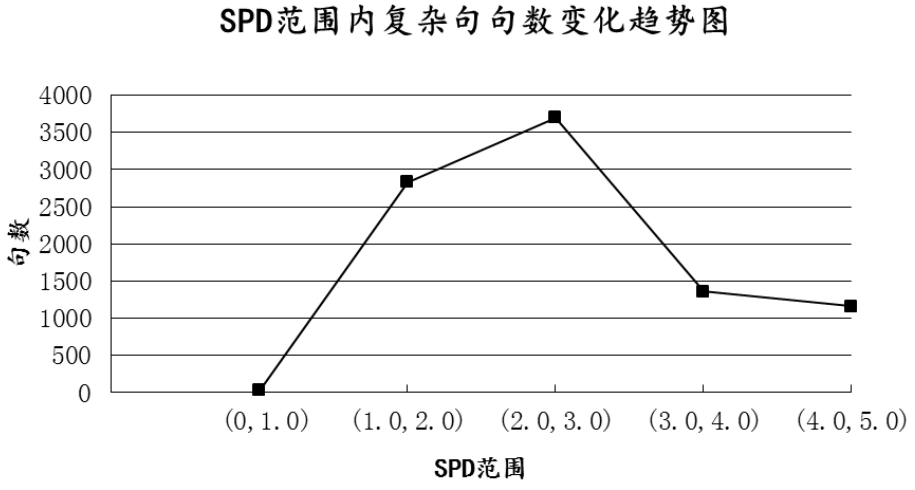


**图 1 APD 范围内复杂句句数变化趋势图**

图 1 说明了 APD 在（2.0,3.0）范围内复杂句的句数最多，体现了复杂句的复杂程度。

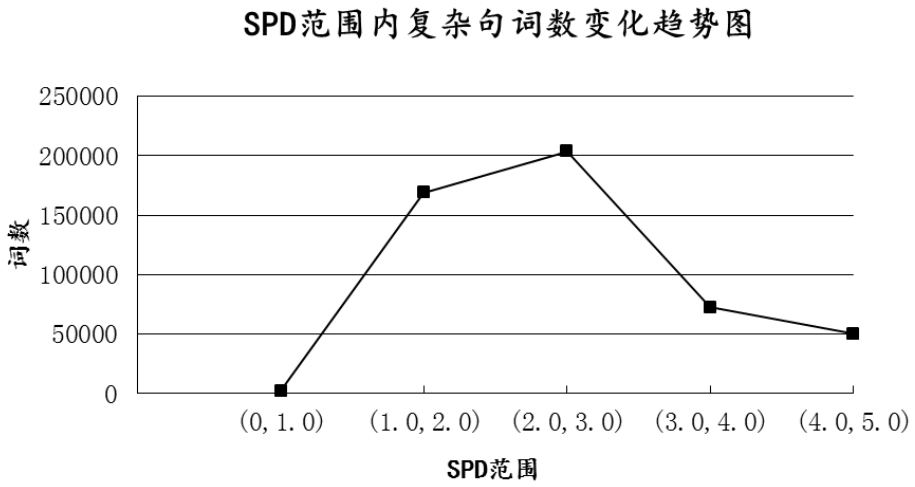


**图 2 APD 范围内复杂句词数数变化趋势图**

图 2 说明了 APD 在（2.0,3.0）范围内复杂句的词数最多，体现了复杂句的复杂程度。

## 3. 从底向上的语料标注方法

从底向上的语料标注方法，采用从底向上、从左到右的层次关系来标注语料。利用句子中的逗号、分号、句号等点号信息，将句子分成若干个小句，分析每个小句的主、状、谓、宾成分，形成了各个 CCC 可能的分析终点位置，然后进行标注。

为了更清楚地描述从底向上的语料标注方法，本文做了如下 5 个解释说明：

（1）标注者：语料校对人员。

（2）校审专家：确定最终标注质量的专家。

（3）CCC Parser：实现自动语料标注的分析器。

CCC Parser 采用移进-规约块分析方法来实现语料标注，输入是经过分词和词性标注的句子，输出是有成分标记和关系标记的句子[12]。CCC Parser 主要关注的是局部语境并且分析速度快，对于简单句来说有更好的分析效果。

CCC 标注示例图 3 所示：

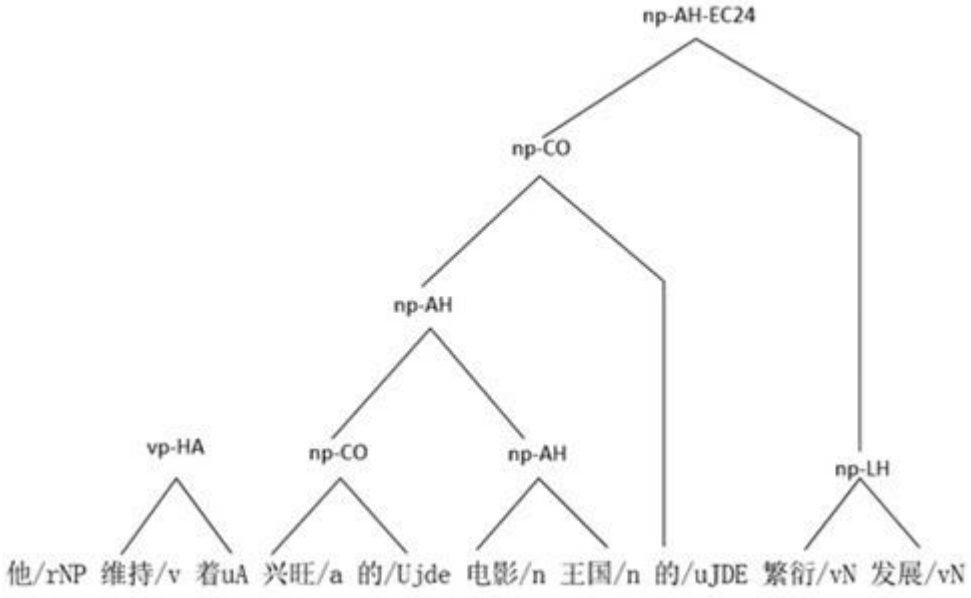


**图 3 CCC 标注示例**

CCC 分析标注的“硬伤”在于 CCC 本身标注错误和对标注句子中小句层面上的特殊从句结构不太准确，需要对其边界和标记进行人工分析与标注[13]。CCC 本身标注错误主要有层次

错误和边界错误。

层次错误是指 CCC Parser 没有正确分析出句子的合理逻辑、合法的句法结构和实际意义而造成的块内成分混乱。

边界错误是指 CCC Parser 没有正确分析出句子主、状、谓、宾成分的整体边界，而是只分析出成分中的一部分，从而造成句义错误。

特殊从句结构主要包括：不同功能词核心控制的补足语，包含典型的主状谓宾结构事件句式片段，增加 EC1 特征标记；并列结构内部的并列成分为特殊的事件句式片段表示，增加 EC1 特征标记；复杂定语从句，不包含助词“的”，增加 EC21 特征标记；复杂定语从句，包含助词“的”，定语部分包含动词谓语、中心语部分是体词块，增加 EC22 特征标记；修饰主体省略的定语从句变形句式，即“的”字结构，增加 EC23 特征标记。

（4）标注手册：制定标注规范的文档。

标注手册主要从基本标注流程、关系标记的分析、成分标记的确定、功能词控制核心的分析处理、CCC 事件内部的标注方法 5 个方面来描述 CCC 标注的准则[14]。它指导标注者如何来校对语料，文档可以直接影响标注的速度和标注质量[15]。

（5）辅助文档：辅助标注者标注的文档。

尽管标注手册已经定义了一些标注的规范，但由于汉语语言的丰富性、复杂性、多样性，对于一些口语化的词语、固定用法、古汉语、省略句等句子信息，在标注的过程中常常会遇一些标注规范中没有涵盖的新的句式、结构。因而在标注的同时，定期举行标注工作总结讨论会，将各个标注者在标注过程中遇到的一些有歧义，在标注规范中没有的一些句式，结构抽取出来，进行讨论，之后将这些有歧义的句式，结构制定一个统一的标注标准。将这个标注标准反馈给校审专家进行审议，审议后将结果返回给标注者，作为新加的标注规范，在以后的标注工作中进行引用。

为了更清楚地描述从底向上的语料标注方法，本文做了如下定义：

**定义 4** 差异率：标注者标注的 CCC 与标准库中的 CCC 相比较正确与否的衡量标准。

差异率=（1－标注正确的 CCC 总数/标准库 CCC 总数）×100%

**定义 5** 一校：利用校对辅助工具，标注者发现并改正 CCC 分析标注“硬伤”。

在一校过程中，标注者要注意发现标注句子中小句层面上的特殊从句结构，对其边界和标记进行人工分析和标注。

**定义 6** 二校：利用 CCC 差异比较工具，发现并确定标注错误。

标注者形成一个环状结构，每个标注随机抽出一校标注语料的 30%给其前面一个标注者进行校对，如果两个标注者的 CCC 标注差异率>2%，则返回一校标注者并重新校对其余 70%的文本。

二校的目的是将标注者在标注过程中出现的一些标注错误以及对歧义句子的标注方法进行统一，把这些歧义句子以及一些标注错误总结为一份标注文档，并将正确标注结果反馈给标注者，避免再犯类似的错误，以提高标注的质量，减少返工。

**定义 7** 校审：确定最终标注质量。

随机抽样所有标注者标注语料的 5%给校审专家审查，如果 CCC 标注错误率>1%，则返回重新校对其余的文本，否则标注质量合格，该部分的语料校对结束。

为了更清晰、直观地描述从底向上的语料标注方法，本文采用流程图的方式来体现从底向上的语料标注方法的整个标注过程，如图 4 所示：

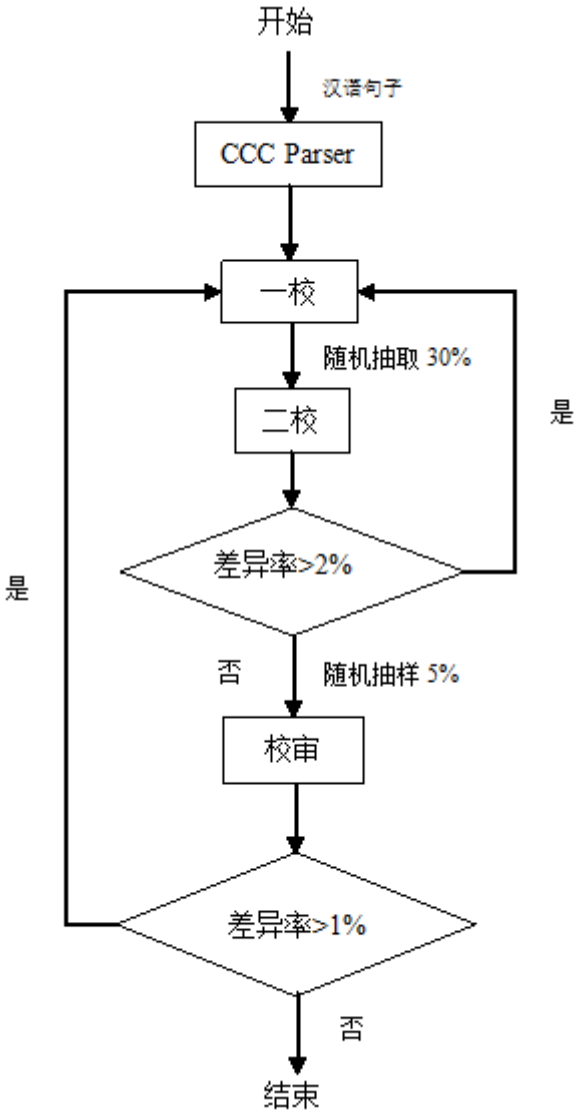


**图 4 从底向上的语料标注方法流程图**

将语料放到 CCC Parser 进行自动分析后，将得到的全部语料给标注者一校，然后标注者随机抽取 30%给另一标注者进行二校，如果二校的差异率小于 2%，则可以送给校审专家校审，否则，一校人员返回校对之前的句子。如果校审的差异率小于 1%，则校对质量合格，该部分语料的校对结束，否则，语料返回到一校人员重新校对。

由于复杂句的句子较长、结构比较复杂等特点，导致在采用从底向上的语料标注方法时标注速度较慢、标注质量不高，所以本文提出了融合从底向上和自顶向下的的语料标注方法。

## 4. 融合从底向上和自顶向下的语料标注方法

首先，利用句子中的逗号、分号、句号等点号信息，采用自顶向下的方法将复杂句子分解为若干小句，再把小句切分成主、状、谓、宾四个成分；然后对切分出来的句子成分按照从底向上的语料标注方法标注；最后，将标注过的句子成分还原成原来的复杂句，再根据小句之间的关系对复杂句采用自顶向下的语料标注方法进行标注。为了更清晰、直观地描述融合从底向上和自顶向下的语料标注方法，增加了如下新的定义：

**定义 8** 块切分辅助工具：将复杂的句子切分成小块的系统工具。

按照句子中的逗号、分号、句号等点号信息，将复杂的句子切分成主、谓、宾、状四个成分，然后提取长度大于等于 2 的块同时去掉重复的块，并记录块的位置信息。由于复杂句的句子比较长，并且句子成分之间的关系比较复杂，经过块切分辅助工具切分之后，就可以

得到简单的块，充分利用 CCC Parser 对块的分析效果更好和分析速度快的优势，得到标注质量高的块。

**定义 9** 块校对：对切分出来的块进行校对。

块的大小相对于复杂句来说，长度较短以及句子成分之间的关系简单，校对速度提高。

**定义 10** 还原：将块还原到原先句子的位置当中。

块校对分析的是块内的关系，而还原之后分析的是块间成分之间的关系，两者属于不同的层次的校对，前者属于块层次，后者属于小句层次。

**定义 11** 句校对：对还原后的句子进行校对。

句校对的目的是通过分析小句之间的关系，发现切分和块校对过程中出现的层次错误、边界错误信息等并及时修改，确保标注质量。

为了更清晰、直观地描述自顶向下的语料标注方法，本文采用流程图的方式来体现融合从底向上和自顶向下的语料标注方法的整个标注过程，如图 5 所示：

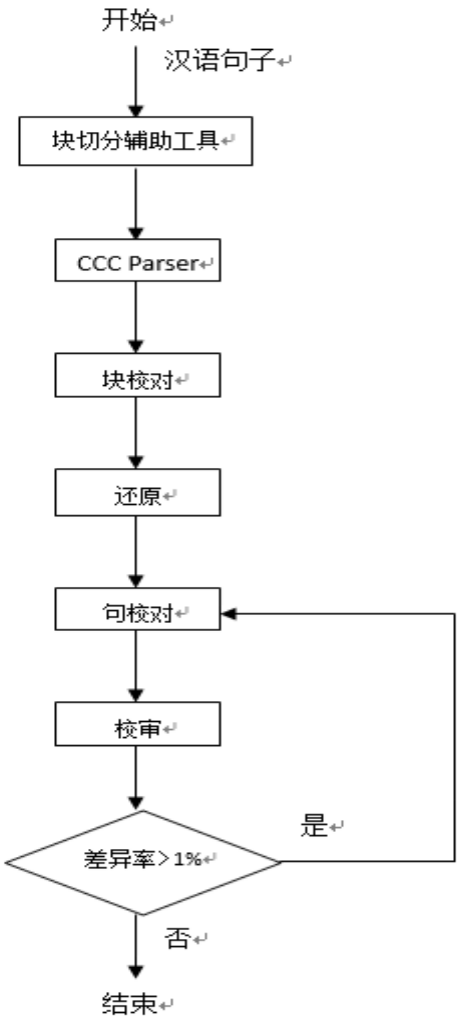


**图 5 融合从底向上和自顶向下的语料标注方法流程图**

利用块切分辅助工具将复杂句切分成块，将块放到 CCC Parser 里进行自动分析，对分析结果实行块校对，块校对完成之后，要将块还原到原来的复杂句中。再对还原后的句子进行句校对。标注者随机抽取 5%给校审专家，如果校审的差异率大于 1%，则语料返回到句校对人员重新校对，否则校对质量合格，该部分语料的校对结束。

## 5. 实验分析

### 5.1 实验数据

目前的标注语料来源主要有三大部分：(1)从现有的汉语句法树库 TCT 中自动提取得到的 CCC 标注库；(2)从北京大学计算语言所标注的人民日报 2000 年语料库按照不同版面提取的语料文件；(3)从山西大学标注的 973 语料库中按照不同体裁提取的语料文件。其中(1)中保留了 TCT 中提取的比较准确的 CCC 标注结果，不需要进行人工校对。(2)、(3)为待标注校对语料，它们的标注规模分别是：(2)200 万词左右；(3)100 万词左右。

## 5.2 实验结果

**表 2　两种标注方法所用时间对比表**

| | bottom-up | bottom-up and top-down |
|---|---|---|
| 100 句 | 约 41h | 约 17h30min |
| 平均每句 | 约 24min36s | 约 10min30s |

表 2 是两种标注方法所用时间的比较。对复杂句来说，采用从底向上语料标注方法标注一句的时间约是融合从底向上和自顶向下方法的 2 倍，说明了融合从底向上和自顶向下的语料标注方法节省了大量的时间，解放了人力、物力。

**表 3　两种标注方法分阶段所用时间对照表**

| bottom-up | | bottom-up and top-down | |
|---|---|---|---|
| Parser 整句 | 约 1h | 切分整句 | 约 45min |
| 一校整句 | 约 30h | Parser 切分块 | 约 45min |
| 二校整句 | 约 10h | 校对块 | 约 11h |
| | | 校对整句 | 约 5h |

表 3 是两种标注方法分阶段所用时间的比较。从表中可以看出，虽然融合从底向上和自顶向下的语料标注方法的步骤多于从底向上的，但是总时间却比从底向上的语料标注方法少了很多，说明融合从底向上和自顶向下的语料标注方法对复杂句来说是可行的。

**表 4　两种方法分阶段差异率**

| bottom-up | | bottom-up and top-down | |
|---|---|---|---|
| Parser--一校 | 40.58% | Parser--块校对 | 27.66% |
| 一校--二校 | 34.16% | | |
| 二校--审校 | 36.03% | 整句校对--审校 | 11.29% |

表 4 是两种方法的差异率的比较。从表中可以看出，虽然从底向上的语料标注经过了两次校对，但是从底向上方法中的 Parser--一校和二校--审校的差异率都比融合从底向上和自顶向下方法中的 Parser--块校对和整句校对--审校的差异率高很多，说明融合从底向上和自顶向下的语料标注方法标注质量更高。

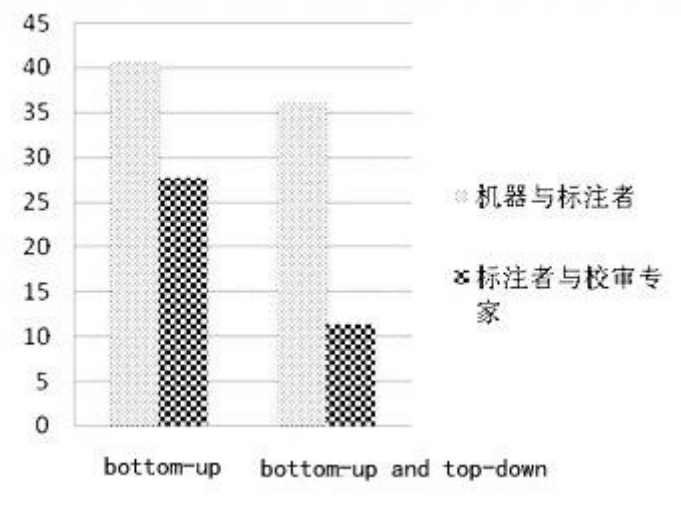


**图 6　机器与标注者、标注者和校审专家的差异率在两种方法中的体现**

图 6 是机器与标注者、标注者和校审专家的差异率在两种方法中的体现。从图中可以看出，自顶向下方法中机器与标注者、标注者与校审专家的差异率都比从底向上的低，说明了自顶向下标注方法的标注质量更好，同时也体现了标注者在语料标注中的重要性。

## 6. 未来与展望

本文通过自顶向下和从底向上的语料标注方法的比较，证明了自顶向下方法对语料标注的有效性，同时也节省了大量的人力、物力，缩短了语料标注时间和提高了语料标注质量。为了更好的提高标注质量，我们在继续改进 parser 算法的同时，将标注质量合格的语料作为 parser 的训练语料，通过训练使 parser 学习到更多的句式结构，这将是以后研究的重点。

## 参考文献


[1] 刘海涛, 赵怿怡. 基于树库的汉语依存句法分析[J]. 模式识别与人工智能, 2009, 22(1): 17-21.

[2] 许建潮. Web 挖掘中若干问题的研究[D]. 长春: 吉林大学, 2005.

[3] 张春祥, 栾博, 高雪瑶, 等. 基于句法分析的汉语词义消歧[J]. 计算机应用研究, 2014, 31(1): 40-42.

[4] Ide N. Annotation Science From Theory to Practice and Use[J]. 2007.

[5] Xue N, Chiou F D, Palmer M. Building a large-scale annotated Chinese corpus[C]//Proceedings of the 19th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 2002: 1-8.

[6] 王东波, 谢靖. 基于清华汉语树库的有标记联合结构统计分析[J]. 现代图书情报技术, 2010, 4: 12-17.

[7] 周强. 汉语基本块描述体系[J]. 中文信息学报, 2007, 21(3): 21-27.

[8] 周强, 赵颖泽. 汉语功能块自动分析[J]. 中文信息学报, 2007, 21(5): 18-24.

[9] Ming Lai E Y, Tan L, Wong V, et al. The OPT-ional Phenomenon in Singapore English: A Corpus-based Approach Using Time Annotated Corpora[J]. Procedia-Social and Behavioral Sciences, 2013, 95: 431-441.

[10] Kim J D, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature[J]. BMC bioinformatics, 2008, 9(1): 10.

[11] 周强，汉语语篇标注库的初始语料准备，清华大学信息技术研究院语音和语言技术中心，技术报告 TH-RIIT-CSLT-TR-20131205

[12] 王大鹏. 基于 UAM 的 Stanford parser 多层级句法标注实例评析[J]. 电子测试, 2013 (9).

[13] 宇航, 周强. 汉语基本块标注系统的内部关系分析[J]. 清华大学学报: 自然科学版, 2009 (10): 1708-1711.

[14] 周强，汉语概念复合块标注规范 (Ver 1.0)，清华大学信息技术研究院语音和语言技术中心，技术报告 TH-RIIT-CSLT-TR-2013-11-01

[15] 史宪军. 文本信息人工标注辅助系统的设计与实现[D]. 太原理工大学, 2008.