平行树库的标注校对模式研究*

梁军，柴玉梅，昝红英，穆玲玲，韩英杰，张坤丽
郑州大学 信息工程学院，河南 郑州 450001

摘要：目前机器翻译采用的方法大多是基于统计或统计与规则相结合的方法，而平行树库对于构建源语言到目标语言的翻译模型有着至关重要的作用。针对此本文对平行树库的标注校对模式进行研究并开发出一个依据宾州树库标准快速构建英汉平行双语句法树库的标注、校对工具。该工具首先使用 Berkeley parser 解析出句子的短语结构树，并提供可视化显示、操作界面；然后用户可以使用自定义约束条件或采用直接拖曳方式对机器生成的句法树进行校对修改并最终得到正确的句法树结构。通过构建 4000 句对中英文平行句法树库，表明该工具可大大提高标注校对效率，并减少人为错误；并根据人工校对的结果对当前句法分析器的主要错误进行分析。

关键词：平行树库；可视化；标注；短语结构

Exploration of Tagging and Verifying on Parallel TreeBank

LIANG Jun, CHAI Yumei, ZAN Hongying, MU Lingling, HAN Yingjie, ZHANG Kunli
School of Information Engineering, Zhengzhou University ,Zhengzhou, Henan 450001, China

Abstract: The current methods are mostly based on statistical machine translation or the method of combining statistics and rules, at the same time parallel treebank has a vital role for building translation model from the source language to the target language. In view of this, the paper study on annotation model of the parallel treebank and develop a tagging and verifying tools for quickly building parallel English-Chinese bilingual treebank based on the standards of Penn TreeBank. The tool firstly parses out the phrase structure tree of sentence by using Berkeley parser and provides a visual interface; then the user can modify and verify the syntax tree and finally get the correct syntax tree by adding custom constraints or directly dragging on the visual syntax tree produced by machine. By building 4000 pairs parallel English-Chinese Treebank, indicating that the tool can greatly improve efficiency labeling proofreading and reduce human error, and we analyze the main error of parser according to the artificial tagging result.

Keywords: Parallel TreeBank; Visual; Tagging; Phrase structure

1 引言

语料库的句法标注是语料库语言学研究的前沿课题[1],[2]，它的目标是对语料文本进行句法分析和标注，形成树库(Treebank)语料。近年来，国内外研究人员在这些方面进行了深入探索，开发完成了许多大规模的树库。在英语方面，有英国的 Lancaster-Leeds 树库[3]和美国的 Penn 树库[4]；德语方面有 NEGRA 树库[5]和 TIGER 树库[6]；捷克语方面，有布拉格依存树库(PDT)[7]；汉语方面，有美国宾州大学的 Penn 中文树库和台北“中研院”的 Sinica 中文树库[8]，北京大学的汉语树库[9]，以及清华大学的 TCT(Tsinghua Chinese Treebank)句法树库[10]。

近年来，随着机器学习方法的不断进步，基于机器学习技术的句法解析也有了不少新的突破，斯坦福大学利用神经网络模型进一步提升了短语、依存句法解析的准确性[11],[12]。但这些工作都需要依托人工建立的树库资源，这些树库作为包含句法结构信息的深加工语言资源，对语言学研究和 NLP 自动句法分析起到了非常重要的基础作用，但上述树库都是针对单语种构建的大规模短语树库，没有涉及双语平行树库的构建。“平行语料库”由英文术语“parallel corpus”翻译而成。对“parallel corpus”的定义主要有三种观点。一种观点认为平行语料库由取样标准相同的单语语料库构成，即将“parallel”理解为语料取样标准的一致性。另一种观点认为平行语料库由原文本及其对应的翻译文本构成，即将“parallel”理解为一种翻译对应关系。而第三种观点比较折中，认为以上两种语料库均属于平行语料库范畴。目前中英文平行语料库有 Babel 语料库，它有句对齐和词性标注，语料取自英文报刊文章及其中译文，主要用于中英语言对比研究[13]；专利语料库，由中国专利的中英文摘要或专利中英文可比较语料挖掘的句对构建，句子结构复杂，但是没有进行深加

*基金项目：本文承国家自然科学基金项目（61402419，61272221）、国家社会科学基金项目（14BYY096）、国家高技术研究发展 863 计划（2012AA011101）、河南省科技厅科技攻关计划项目（132102210407）,河南省科技厅基础研究项目（142300410231，142300410308）、河南省教育厅科学技术研究重点项目（12B520055，13B520381）、国家重点基础研究发展计划 973 课题（2014CB340504）、河南省高等学校重点科研项目（15A520098）支持资助。

工[14]。大连理工大学曹井香，黄德根等人[15]提出了面向机器翻译研究的融合短语结构和依存分析的短语依存树，并构建了大连理工大学平行 PDT。但这些平行树库的标注构建均采用人工标注校对，太依赖于标注人员的语言学经验。Zhang K[16]等人也在构建双语树库进行了尝试，本文根据现有的标注体系框架对平行树库的标注校对模式进行研究，希望在进行人工标注的时候可以充分利用现有语料的统计信息以避免人为错误，利用该标注模式尝试建设一个小规模的中英文双语平行树库，并为此设计并实现了一个高效的标注校对工具[1]。

2 标注系统概述

在英语方面，美国的 Penn 树库的标注体系经历了一个从简单到复杂的不断进化发展过程。最初的 PTB-1[4]采用了骨架分析(Skelton Parsing)思想，形成比较扁平的句法结构树。随后，在扩充版本(PTB-2)[17]中，增加了一些功能标记，用于标注句子中主要句法成分的语法功能，希望能据此自动抽取出句子的谓词-论元(Predicate-Argument)信息。从 2002 年起，他们进一步提出了命题库(PropBank)构建计划[18]，在 PTB-2 上明确标注句子中各个动词的谓词-论元信息，希望借此建立从句法到语义的重要桥梁。捷克的 PDT 项目[7]则设计了三个层次的标注信息：词法、句法和语义。在词法层面上，充分利用了捷克语丰富的形态变化信息，总结了 4200 多个词类标记，在此基础上形成的句法依存树。在德国的 TIGER 树库中，研究人员采用了一种层次结构和依存关系相结合的标注体系：底层的句法成分主要采用层次结构，可以保留大量丰富的描述信息；高层的语法关系则采用依存结构，描述句子中各主要成分与中心动词之间的各种句法依存关系，形成一种功能强大、处理灵活的描述体系，特别适合于像德语那样语序比较自由的语言。

在汉语方面，美国宾州大学于 1998-2000 年以新华通讯社的 10 万词新闻文本为语料，率先建成了宾州中文树库 CTB-I。2003 年进一步完成了 CTB-II 的标注，语料内容增加了人民日报、香港新闻电讯和从其他语言翻译过来的中文稿件，规模扩大到 40 万词。最新版本的 CTB8.0[2]增加了新的通讯社、杂志文章以及政府文件，已经达到 162 万词级别。宾州树库对中英文标注基本采用相同的 PTB-2 标注体系。对于英文设计了 36 种成分标记[19]，26 种短语结构标记[20]，句子则只能由 S，SBAR，SBARQ，SINV，SQ 引导；对于中文则设计了 34 种词性标记[21]，17 种短语结构标记[22]，句子由 IP，CP 引导，标注体系的详细信息可查看附录。国内对中文树库的构建也有较为深入的研究，主要有北京大学构建的汉语句法树库[3]和清华大学构建的 TCT 树库[4]。这两个汉语句法树库一脉相承，均以汉语传统的层次分析法为理论基础标注句子层次，采用相对较小的词类标注集，并在词类标注的基础上对直接成分之间的句法关系进行了标注。

由于本文尝试构建的小规模中英文双语平行树库包含中文树库和英文树库，故需要采用一种通用的标注系统，鉴于宾州树库标注系统涵盖中文、英文，因此选用宾州树库标准。同时，在进行标注的时候也参考了伯克利大学对中文句法分析进行的分析[23]。

3 标注校对模式的研究

现有标注模式大多以单一语种为研究对象，而不是从构建多语种平行树库的角度，故存在以下几个问题：

1) 标注语料选取仅考虑某一种语言，语料树库构建完成后不能为跨语言的工作提供依据和帮助；
2) 在标注语料时更多地依靠标注者的语言学经验，往往没有参考语料库的统计规律和特征；
3) 由于大多数语料库是单语种语料库，在构建过程时不会考虑其他语种的语法语义结

[1] 依托于本文对标注校对模式的研究开发出的校对工具已开源：https://syntree.github.io/

[2] https://catalog.ldc.upenn.edu/LDC2013T21

[3] http://ccl.pku.edu.cn:8080/WebTreebank/WebTreebank_Readme.html

[4] http://cslt.riit.tsinghua.edu.cn/~qzhou/chs/Resources.htm

构。

3.1 标注校对模式简介

本文根据以上在现有语料树库构建中出现的问题，对双语语料库构建模式进行探索尝试并基于此开发了一个平行树库的标注校对工具，在构建树库时采用人工校对加机器辅助的方式进行标注校对工作。具体的标注校对流程如图 1 所示。

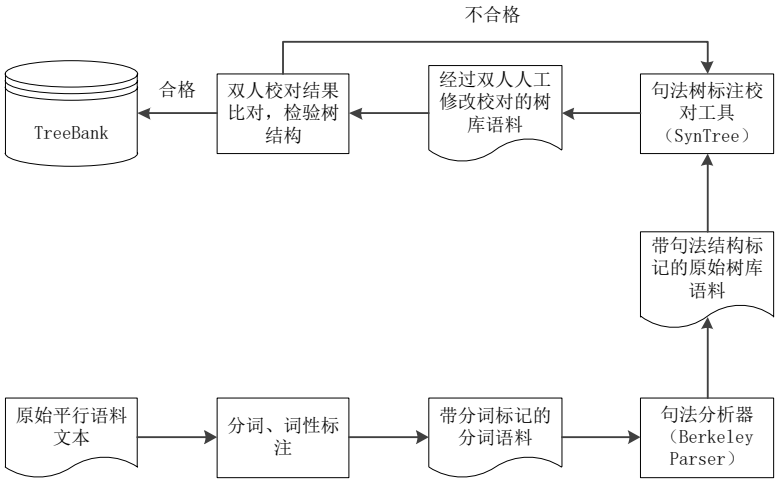

图 1 双语平行树库加工流程示意图

在收集得到双语平行语料之后，首先对原始句子进行分词和词性标注处理，得到带词性标记的分词语料后交给人工校对，然后利用伯克利句法分析工具对分词和词性标注基本正确的语料进行自动的句法分析并输出带有句法结构标记的初始树库 T_0；再在 SynTree 句法校对工具的辅助下，由双人对初始树库进行校对，修正其中的错误，得到经过人工校对的树库语料 T_1、T_2；然后对 T_1 和 T_2 的一致率进行验证，对不符合要求则重新校对，直到一致率达到标准为止。

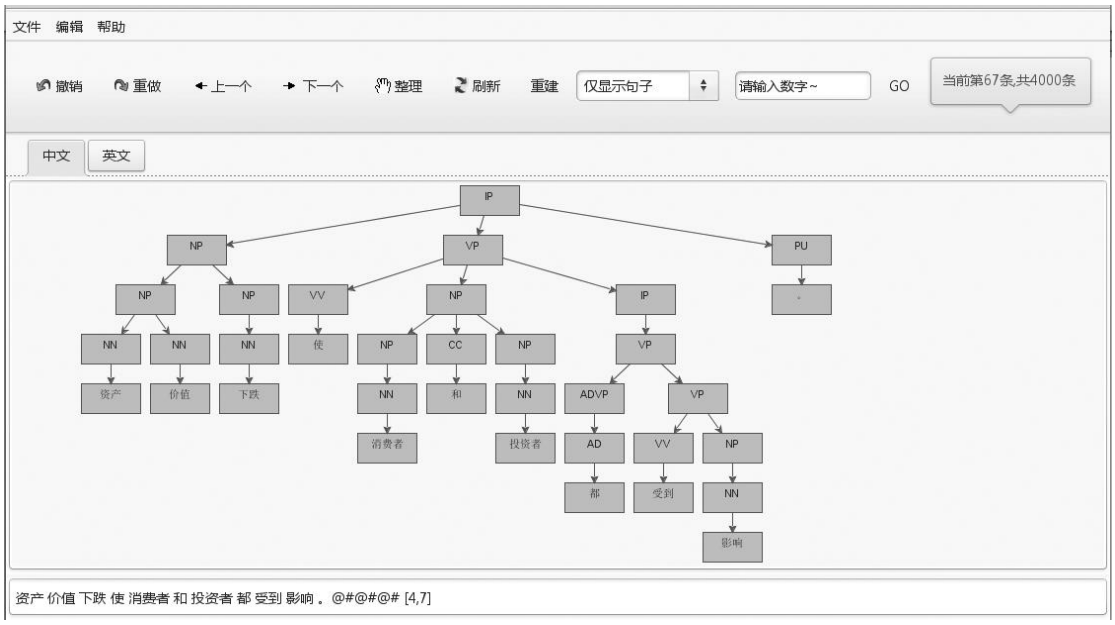

图 2 标注校对工具主界面

根据宾州树库的标注体系本文构建了一个针对双语平行树库标注校对的系统，标注工具的界面如图 2 所示。该系统集成了伯克利句法解析工具[5]，并在该句法解析工具的基础上进行修改使得可以通过添加约束条件让机器参照约束条件自动生成句法解析结果（该功能将在

[5]https://code.google.com/p/berkeleyparser/

3.2 节进行详细描述），同时也提供中英文双语句法树同时校对的功能（鉴于中英语句法树图形化之后占据空间较大，因此分为两个窗口显示，但是中英文句法树展示窗口显示的是相对应的句对的句法树），可以方便的进行平行句法树的调整校对工作。

3.2 词语切割的标注模式

原始平行语料包含中文和英文，对于中文语料需要先进行分词处理，而英文语料则不需要，得到分词工具输出的带分词标记的分词语料后可以作为句法分析器的输入，从而自动得到带句法结构标记的原始树库语料。当然，由分词工具得到的分词结果一定不是 100%正确的，在 SynTree 校对工具中可以对分词错误的词语进行调整，然后点击工具栏重建按钮即可得到修正分词结果后对用的树库，如图 3 所示，是校正分词结果后的句法解析结果对比。

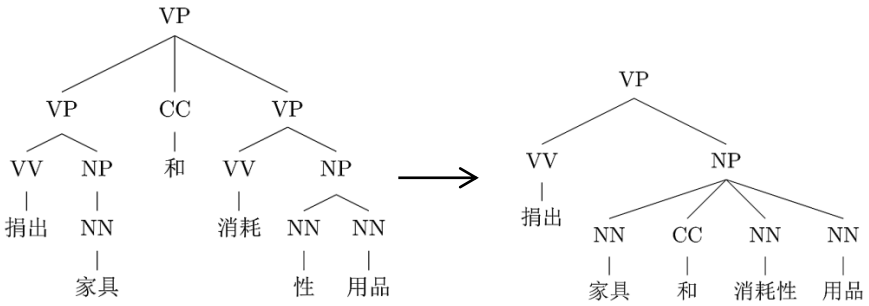

图 3 分词结果对句法树的影响

由上图可以看出分词结果对句法解析工具生成的结果会产生很大的影响，因此，为了保证句法解析工具解析生成句法树的正确率，首先要确保分词的正确性。在 SynTree 校对工具下方的句子文本显示窗口，可以直接对分词结果进行编辑，然后使用重建功能重新生成句子相应的句法树。

3.3 短语结构的校对模式

得到带句法标记的原始树库之后，就可以使用 SynTree 校对工具对句法树结构进行调整，这里提供两种调整模式：①约束条件调整，②拖曳调整。约束条件调整是通过人为添加一些句子的限定条件来让句法解析器参照这些约束生成句法树，这样可以起到调整句法树结构的目的，同时又可以防止人为的主观错误。拖曳调整这是直接通过拖曳句子树结点以达到修改句法树结构的目的，通过这种方式可以将句法树任意修改成符合人的主观意愿的结构，但是可能造成句法结构的不合理。因此这种调整方式作为“约束条件调整”方式的辅助修改方式，当无法通过添加约束条件生成正确的句法结构时，可以通过直接拖曳的方式对句法树进行调整。本节主要介绍如何通过约束条件对句法树结构进行调整，拖曳调整的方式将在第 4 节作为辅助功能进行介绍。

为了方便给句法树添加自定义约束条件，本文为句法树重新定义了几个新的概念：叶子节点的直接父节点、叶子节点的深度和叶子节点间的耦合度。

直接父节点（DP）：直接父节点是指该节点有且仅有一个子节点，如图 4 中 ADVP、AD、PU……都是直接父节点，例如 VV 是节点 3（“使”）的直接父节点，NP、NN 是节点 4（“消费者”）的直接父节点，那么可以表示为：

$$DP(3) = \{ VV \}; DP(4) = \{ NP, NN \}$$

对于节点 4 可以看到有两个直接父节点 NP 和 NN，其中直接父节点 NP 称为节点 2 的最远直接父节点，记作：FDP(4) = { NP }。

叶子节点的深度（Deepth）：通常情况下，叶子节点的深度是根节点到叶子节点的距离，本文重新定义为根节点到该叶子节点最远父节点的距离。例如，节点 3 的最远直接父节点是“VV”，那么 Deepth(3) = 2；对于节点 6（“投资者”），其最远直接父节点是“NP”，那么 Deepth(6) = 5。

叶子节点间的耦合度（Coup）：两个节点间的耦合度是指两个叶子节点之间的关联关系，

两个叶子节点的关联关系越紧密则它们之间的耦合度值[6]Coup(x, y)就越小，如图 4 中 Coup(3, 4) < Coup(3, 5)。耦合度值可以按照下面的公式进行计算：

$$Coup(x, y) = |Deepth(x) – Deepth(y)| + 1$$

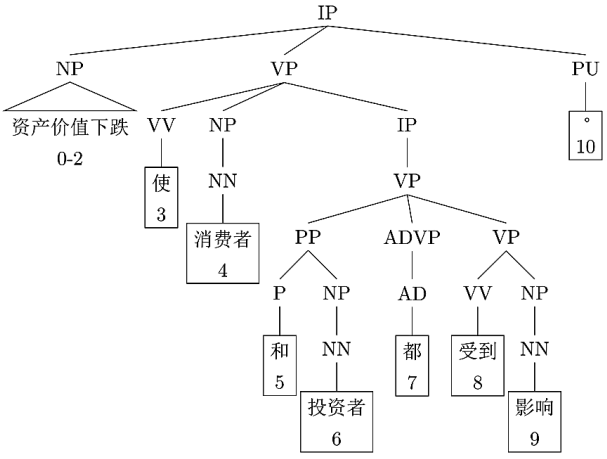

图 4 原始语法树

根据上面的定义，可以很简单的计算出来节点 4 和节点 6 之间的耦合度，Coup(4, 6)= |Deepth(4) –Deepth(6) | + 1 = |2 - 5| + 1= 4，而 Coup(3, 4) = |2 - 2| + 1 = 1，那么就有 Coup(3, 4) < Coup(4, 6)。但是根据语义理解，节点 4 和节点 6 的关联度应该比节点 3 和节点 4 的关联度小，即 Coup(3, 4) > Coup(4, 6)，为了修正这个错误，就需要在使用句法解析工具进行自动句法生成时添加额外的限定条件。本文实现的句法校对工具设计并实现了该功能，在将带解析的句子传递给句法解析工具时只需要在后面添加上约束条件即可使得句法解析工具在约束条件下去进行句法分析。如图 4 所示，本来该句子解析时的输入为：“资产价值下跌使消费者和投资者都受到影响。”，添加约束条件后变为“资产价值下跌使消费者和投资者都受到影响。@#@#@# [4,7]”。其中“@#@#@#”作为句子和约束的分隔符，“[4, 7]”是约束条件（句子的词语从 0 可是编号，标号符号按一个词语处理），表示节点 4、5、6 有较强的关联关系。添加约束条件后，使用重建功能，得到新的句法解析树如图 5 所示：

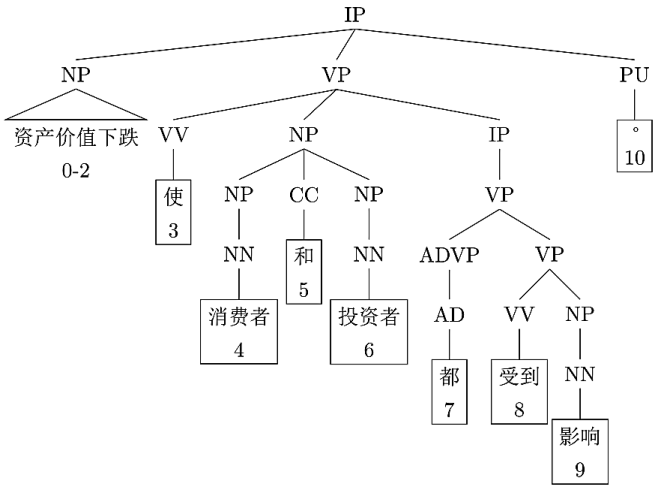

图 5 添加约束后的句法树

采用这种添加约束的方式，也可以对单个词语的词性进行约束，加入节点“和”的词性应该是“ETC”，那么就可以通过添加约束“@#@#@# ETC[5,6]”来限定词“和”的词性标记。

[6]根据耦合度值的定义，可以知道其满足交换性，即 Coup(x, y) = Coup(y, x)。

3.4 双语平行校对模式

由于本文试图探索双语树库的构建，在收集语料时，每个句子都是中英文对照的。如图6所显示的正是句子“资产价值下跌使消费者和投资者都受到影响。（The fall in asset values affected consumers and investors .）”的中英文树库的对应关系。在工具中分别有中、英文校对界面，可以分别显示中英对照的句子的相应句法树。

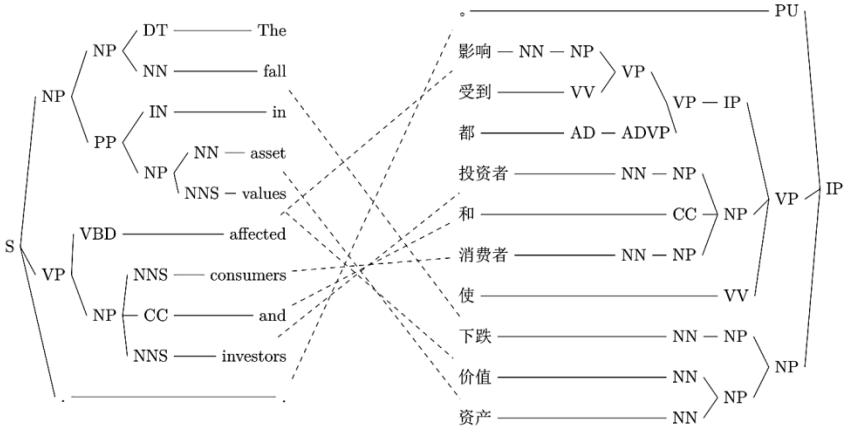

图 6 中英文句法树对应关系

对于英文的校对方式可以参照中文短语结构的校对方法进行，因为英文的语言特性没有分词的需要。

4 辅助功能

基于约束条件的短语树库校对模式从总体上保证了树库标注的准确性，提高了校对的效率，而增加便捷的辅助功能以优化人机界面，既能方便标注人员操作，提高树库标注效率，也可对树库构建的一致性起到一定的促进作用。本系统实现树库校对的基本功能的基础上，设计实现了一系列辅助树库建设的外围功能，比如拖曳调整树库、整理树结构、撤销/重复等。

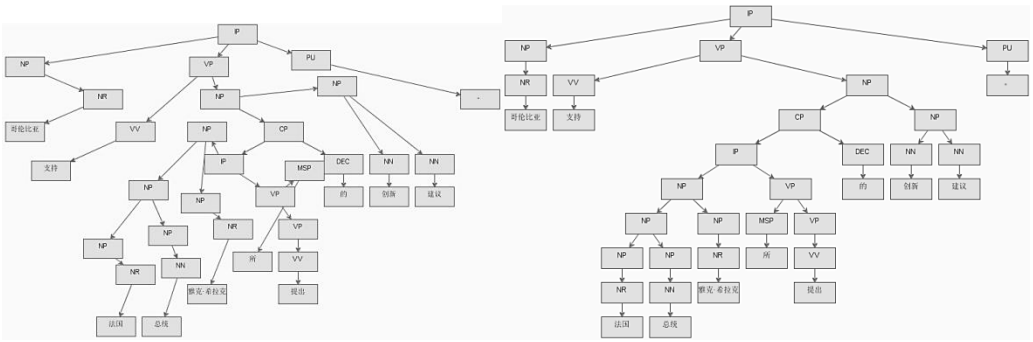

图 a 通过拖曳调整得到的句法树　　图 b 整理后的句法树

图 7 SynTree 工具的辅助功能

使用图形界面操作是，对树结构进行拖曳操作是一种非常便捷、友好的方式，本系统也设计实现了通过拖曳方式调整树库的功能。使用本系统可以对树的任意节点进行删除、修改和增加节点，同时也可以对节点间的连接关系进行调整。但是由于该方法完全凭借人对句法树的认识去操作，容易造成结构的不合理，树库构建的不一致性，因此该方法虽然简单易用但仅仅作为对句法树校对的一个辅助方法，并不推荐使用。

通过拖曳操作调整树库时，还有一个问题就是造成树结构的层次错乱如图 7a 所示，因此，本系统特意设计实现一个整理句法树的功能，在拖曳句法树节点造成树结构层次不清晰时可以通过点击工具栏的“整理”按钮来对句法树进行一键整理得到结构层次清楚的句法树，

如图 7b 所示。

此外，考虑到标注过程中偶尔会有一定的误操作，为了能够让标注人员在不用重新标注的情况下就能快速地修正错误，系统设计并实现了标注操作的撤销与重复功能。标注人员可通过工具栏的“撤销”和“重做”按钮执行。

5 双语树库标注实践分析

在对 4000 对中英文对照语料标注完成后，本文对提出的校对模式对一致率的贡献做出的统计分析，同时对比句法分析器生成的句法分析结果和人工标注的句法分析结果对当前句法分析器的错误进行分析统计。

5.1 一致率统计分析

为了检验标注模式及实现工具的效果，并在标注过程中发掘更加合理高效的标注模式和机器辅助算法，本文将标注工具应用于中英文双语平行短语结构树库构建工程。语料是抽取 CWMT 汉英新闻训练语料中的 4000 句对中英文平行语料，中文经过分词处理后与英文一起作为标注的原始预料。每个句对都采用双人校对的方式，并由对结果进行比对，一致的接受加入语料库，不一致的由第三人评判验收。为验证工具的有效性，随机选取 2000 对不使用工具进行标注，另外 2000 对使用 SynTree 工具进行标注。

表 1 双人标注一致率统计

	中文			英文		
	一致	不一致	一致率	一致	不一致	一致率
未使用工具标注	236	1764	11.8%	704	1296	35.2%
使用SynTree标注	694	1306	**34.7%**	1268	732	**63.4%**

根据表1统计情况可以看出通过使用该工具，双人校对一致率有较大提高，中文单句一致率提高到34.7%，英文单句一致率提高到63.4%（该句子的句法树完全一致才认为一致，而不是按照句法树的子树进行统计）。

5.2 中英文句法树错误统计分析

在人工标注完成双语句法树后，本文采用 Kummerfeld et al.等人[18]的分析工具对中英文句法树的错误进行分析，其中中文句法树均约有 3959 句，英文句法树有 3984 句。

表 2 中文句法分析错误统计[7]

错误类型	数量	占比
UNSET add	2536	9.59%
Verb taking wrong arguments	2264	8.56%
UNSET remove	1724	6.52%
Unary - IP over VP	1558	5.89%
Modifier Attachment	1416	5.35%
Different label	1090	4.12%
NP Internal - add	990	3.74%
Co-ordination	963	3.64%
Noun boundary error	928	3.51%
Clause Attachment	820	2.95%
UNSET move	779	2.95%
NP Internal - remove	571	2.16%
NP Internal - UNSET add	567	2.14%
Single Word Phrase - remove	565	2.14%
Others	9673	36.57%

上表中及下表中的错误类型具体含义详见 Kummerfeld et al. 2013 年 ACL 论文，具体见参考文献 15。从上表可以看出 Verb taking wrong argument（动词范围识别错误）、Modifier

[7] 错误分析中的错误类型详见 Kummerfeld et al.2013 年 ACL 论文，具体见参考文献 15.

Attachment（修饰关系错误）、Noun boundary error（名词短语范围识别错误）在句法分析中占很大比重，这表明在句法分析中词语语义边界识别仍然是一个比较难以由机器解决的问题。同时也可以看出 NP Internal（名词短语结果错误）也是一个比较严重的问题，在中文中名词短语内部的修饰结果关系也是一个具有挑战的任务。

表 3 英文句法分析错误分析

错误类型	数量	占比
PP Attachment	1401	13.20%
UNSET add	1062	10.05%
Modifier Attachment	980	9.27%
Different label	848	8.02%
UNSET move	834	7.89%
UNSET remove	667	6.31%
NP Internal - Single Word Phrase - relabel	456	4.31%
Co-ordination	305	2.89%
Single Word Phrase - remove	251	2.37%
Single Word Phrase - remove - for guidelines	247	2.34%
Single Word Phrase - add	243	2.23%
Co-ordination - Wrong pair	213	2.01%
Others	3064	29.0%

从上表可以看出在英文句法分析中，PP Attachment（介词短语的修饰对象）以及 Modifier Attachment（修改关系错误）也是较为严重的问题，而在英文中动词的修饰范围识别则没有中文中那么严重的问题，在统计中仅出现 156 例错误远低于中文的 2264 例错误。

6 结语

本文根据人工构建双语平行树库中存在的误标、效率低等问题，构建了一个平行树库的标注校对工具，在实现过程中中英文采用统一标准为双语树库的进一步探索提供了坚实的基础。经过实践检验，该工具快捷高效，可以极大提高人工标注效率，降低人工错标率。当然平行语料库的构建除了句子对齐之外，更重要的是短语或词的对齐，之后的工作也将对此作出进一步探索。

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] Yıldız O T, Solak E, Görgün O, et al. Constructing a Turkish-English parallel treebank[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014: 112-117.

[3] Leech G, Garside R. Running a grammar factory: the production of syntactically analysed corpora or treebanks[J]. Johansson and Stenströ. 1991: 15-32.

[4] Marcus M P, Marcinkiewicz M A, Santorini B. Building a large annotated corpus of English: The Penn Treebank[J]. Computational linguistics. 1993, 19(2): 313-330.

[5] Skut W, Brants T, Krenn B, et al. A linguistically interpreted corpus of German newspaper text[J]. arXiv preprint cmp-lg/9807008. 1998.

[6] Brants S, Hansen S. Developments in the TIGER Annotation Scheme and their Realization in the Corpus.[Z]. 2002.

[7] Hajic J. Building a syntactically annotated corpus: The prague dependency treebank[J]. Issues of

valency and meaning. 1998: 106-132.
[8] Huang C, Chen F, Chen K, et al. Sinica Treebank: design criteria, annotation guidelines, and on-line interface[Z]. Association for Computational Linguistics, 2000:29-37.
[9] 周强，张伟，俞士汶. 汉语树库的构建[J]. 中文信息学报. 1997(04).
[10] 周强. 汉语句法树库标注体系[J]. 中文信息学报. 2004(04).
[11] Socher R, Bauer J, Manning C D, et al. Parsing with compositional vector grammars[C]//In Proceedings of the ACL conference. 2013.
[12] Chen D, Manning C D. A fast and accurate dependency parser using neural networks[C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 740-750.
[13] Xiao R．The Babel English-Chinese Parallel Corpus [DB/OL]. [2013-2]. http://www.lancaster.ac.uk/fass/projects/corpus/babel/babel.htm
[14] Lu B, Tsou B K, Jiang T, et al. Mining large-scale parallel corpora from multilingual patents: An English-Chinese example and its application to SMT[C]//Proceedings of the First CIPS-SIGHAN Joint Conference on Chinese Language Processing. 2010: 79-86.
[15] 曹井香, 黄德根, 王伟, 等. 中英平行短语依存树库构建[J]. 大连理工大学学报, 2014, 54(1): 91-99.
[16] Zhang K, Zan H, Han Y, et al. Preliminary Study on the Construction of Bilingual Phrase Structure Treebank[M]//Chinese Lexical Semantics. Springer International Publishing, 2014: 403-413.
[17] Marcus M, Kim G, Marcinkiewicz M A, et al. The Penn Treebank: annotating predicate argument structure[Z]. Association for Computational Linguistics, 1994:114-119.
[18] Kingsbury P, Palmer M, Marcus M. Adding semantic annotation to the penn treebank[Z]. Citeseer, 2002252-256.
[19] Santorini B. Part-of-speech tagging guidelines for the Penn Treebank Project (3rd revision)[J]. 1990.
[20] Bies A, Ferguson M, Katz K, et al. Bracketing guidelines for Treebank II style Penn Treebank project[J]. University of Pennsylvania. 1995, 97.
[21] Xia F. The segmentation guidelines for the Penn Chinese Treebank (3.0)[J]. 2000.
[22] Xue N, Xia F, Huang S, et al. The bracketing guidelines for the Penn Chinese Treebank (3.0)[J]. 2000.
[23] Kummerfeld J K, Tse D, Curran J R, et al. An Empirical Examination of Challenges in Chinese Parsing[C]//ACL (2). 2013: 98-103.