基于连接依存树的汉语篇章结构分析平台 *

李艳翠[1,2]，孙静[2]，冯文贺[3]，周国栋[2*]

（1. 河南科技学院信息工程学院，河南 新乡 453003；2. 苏州大学计算机科学与技术学院，江苏 苏州 215006；3. 河南科技学院文法学院，河南 新乡 453003）

摘要：该文采用基于连接依存树表示体系的汉语篇章结构语料构建汉语篇章结构分析平台。该语料标注内容包含子句、连接词、篇章关系、篇章单位主次和篇章结构树等。在此语料上，采用自底向上的方法进行汉语篇章结构分析，包含子句识别、连接词识别与分类、篇章关系识别、篇章单位主次识别和篇章结构树构建等子任务。最后给出了各个子任务的实验结果及汉语篇章结构分析平台的整体性能。

关键词：连接依存树；篇章结构分析；子句切分；关系识别；

中图分类号：TP391　　**文献标识码**：A

The Platform of Chinese Discourse Structure Analysis based on Connective-driven Dependency Tree

LI Yancui[1, 2], SUN Jing[2], FENG Wenhe[3], ZHOU Guodong[2*]

(1. School of Information Engineering, Henan Institute of Science and Technology, Xinxiang, Henan 453003, China; 2. School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu 215006, China; 3. College of Humanities, Henan Institute of Science and Technology, Xinxiang 453003, China)

Abstract: We give a platform of Chinese discourse structure analysis based on Connective-driven Dependency Tree representation corpus. The corpus is based on connective-driven dependence tree representation system, which contains clause, connective, discourse relation, discourse structure et al. Based on this corpus, this paper presents a Chinese discourse parser, which consists of EDU segmentation, connective recognition, implicit relation recognition, discourse center recognition and discourse structure construction. Finally, the experimental results of each sub task and the overall performance of the Chinese discourse structure analysis are given.

Key words: Connective-driven Dependency Tree；Discourse Structure Analysis；Clause Segmentation；Relation recognition；

1 引言

篇章结构分析是自然语言处理的挑战性课题，相对于词法和句法分析，篇章结构分析研究进展仍比较缓慢，受限于汉语篇章结构语料，目前还没有完整的汉语篇章结构分析平台。文献[1]标注了一定规模的汉语篇章结构语料库，本文拟采用文献[1]所标语料进行汉语篇章结构分析平台构建。下面分别从语料资源和篇章分析两方面介绍汉语篇章结构分析研究现状。

现有英语篇章结构资源主要有修辞结构理论篇章树库(RSTDT)[2]和宾州篇章树库(PDTB)[3]。相比英语，汉语篇章结构研究刚刚起步。资源构建主要采用以下三种方法：1)基于修辞结构理论(RST)的研究。乐明[4]以 RST 为指导，参考汉语复句和句群理论，进行了篇章结构标注的尝试，陈莉萍[5]试图采用 RST 标注汉语篇章。2)基于 PDTB 体系的研究。代表性工作有：Xue[6]和 Zhou 等[7]尝试使用 PDTB 体系标注汉语，Huang 和 Chen[8]从 Sinica Treebank 3.1 中随机抽取了 81 篇石油和旅游领域文档进行标注，完成了 3081 个句对的小规模的中文篇章树库。Zhou 等[9]采用 PDTB 的方法标注了显式句内篇章连接词的论元和关系。张牧宇等[10]在英文篇章关系研究的基础上分析了中英文的差异，总结了中文篇章语义分析的特点，提出一套面向中文的层次化篇章关系体系，发布了哈工大中文篇章关系语料(HIT-CDTB)，目前标注了 525 篇来源于 OntoNotes 4.0 的文本。3)基于汉语本土理论的研究。

收稿日期：2015-6-15 **定稿日期**：2015-8-8

基金项目：国家自然科学基金面上项目（61273320）；河南省教育厅科学技术研究重点项目(14A520080)

参考邢福义 [11]的汉语复句研究成果，华中师范大学标注了汉语复句语料库，该语料库仅关注复句内部关系，没有涉及句子及其以上篇章单位的结构问题。清华汉语树库(Tsinghua Chinese Treebank,TCT)[12]是 100 万汉字规模的汉语句法树库，TCT 中标出了复句内各分句之间的关系信息。但清华汉语树库中没有标注特定复句关系所对应的复句关系词，也没有标注句子之间的关系。

本文所用的汉语篇章结构资源综合了 RST 和 PDTB 的优点并充分考虑了汉语特点，语料标注采用基于连接依存树的汉语篇章结构表示体系，该连接依存树的叶子节点为子句，内部节点为连接词，其中，连接词通过其层级地位表示篇章结构的层次，通过其语义表示篇章关系，另外，连接词所连接的篇章单位根据篇章整体意图区分主次。语料的简单说明见第 2 节，详细内容参看文献[1]。

目前篇章结构分析系统研究主要针对英语，汉语研究由于资源限制进展缓慢。Xue 等 [13]和 Yang 等 [14]基于逗号的篇章分析方法切分汉语句子。其所用语料是根据句法模式自动抽取的，并非基于标注篇章结构语料，不能准确反映实际情况。采用文献[1]所标语料，李艳翠等 [15-17]在进行了子句识别和连接词识别分类的相关工作，孙静等 [18]进行了隐式篇章关系识别的工作。张牧宇等 [19]在哈工大中文篇章关系语料(HIT-CDTB)上进行显式篇章句间关系和隐式篇章句间关系识别，并给出了初步的实验结果。涂眉等 [20]在 TCT 上进行了基于最大熵的汉语篇章结构自动分析方法实验。可见，目前的工作主要是在具体语料上针对汉语篇章结构分析中的某个子任务，还没有完整的汉语篇章结构分析平台。

2 基于连接依存树的汉语篇章结构语料库

2.1 连接依存树

针对汉语篇章结构的一般特点，文献[1]结合 RST 和 PDTB 的优点，吸取 RST 的树形结构、篇章单位主次思想，PDTB 的连接词处理方法，参考汉语复句和句群理论的关系分类等研究成果，采用一种连接依存树(Connective-driven Dependency Tree, CDT)的形式表示汉语的篇章结构。图 1 给出了例 1 的 CDT 表示。

例1 a 浦东开发开放是一项振兴上海，建设现代化经济、贸易、金融中心的跨世纪工程，‖b **因此**大量出现的是以前不曾遇到过的新情况、新问题。| c **对此**，浦东**不是**简单的采取“干一段时间，等积累了经验以后再制定法规条例”的做法，||| d **而是**借鉴发达国家和深圳等特区的经验教训，||||| e<并>聘请国内外有关专家学者，||||| f<并>积极、及时地制定和推出法规性文件，|||| g **使**这些经济活动一出现就被纳入法制轨道。‖ h<例如>去年初浦东新区诞生的中国第一家医疗机构药品采购服务中心，**正因为**一开始就比较规范，||| i<虽然>运转至今，||||| j<并>成交药品一亿多元，|||| k<却>没有发现一例回扣。（chtb_0001）

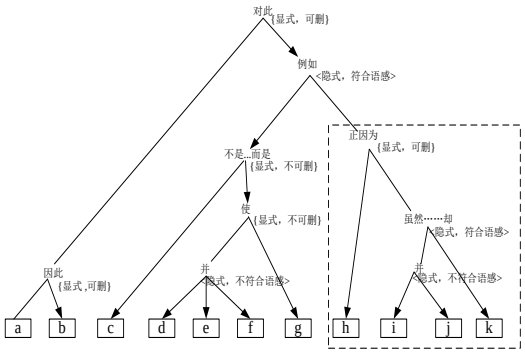

图 1 例 1 的连接依存树实例

图 1 中字母所标记的叶子节点表示基本篇章单位（子句），中间节点表示连接词，各基本篇章单位通过连接词组合后形成高一级篇章单位，进而通过再组合形成更高一级篇章单位，如此层层组合，最后形成一棵篇章结构树。从图 1 可知，连接词的层级地位可以反映篇章结构，连接词本身可以表示篇章关系，连接词所连接的篇章单位有主次之分，通过连线的指向性可区分篇章单位的主次地位。

连接依存树主要包含子句、连接词、篇章结构关系。CDT 的子句含传统单句及复句中的分句，至少包含一个谓语部分，表达一个命题，子句间一定有标点分割。CDT 的篇章连接词不限于现代汉语中的连词，只要对句子和语段起连接作用，能表示句子之间或子句之间关系的语言单位均可称为连接词。篇章结构是一种层次化的树形结构，其中叶子节点为子句，连接词居于不同层级的中间节点。直观上篇章结构分析可看成是各个连接词的不同

层级地位的分析，本质上连接词的不同层级地位反映的是篇章单位的组合层级。结构分析是篇章分析的重要任务，连接词的层级、篇章关系及篇章单位主次地位等都依赖于篇章层级结构的确定。对于篇章关系表示，一般的做法是直接给出并列、转折、因果等抽象关系类型。基于连接依存树的汉语篇章结构表示体系并不直接在树形图中给出这种抽象关系，而是用连接词直接表示篇章单位间的关系。篇章关系体系的构建基于连接词进行不同程度的抽象概括而成。图 2 给出了初步拟定的篇章关系体系，此关系体系在构建时借鉴了汉语复句、汉语句群、修辞结构理论和 PDTB 体系的理论成果。根据子句间的意义关系分类，连接词分为因果类、转折类、并列类和解说类四大类，每一类内部又细分为不同的关系类型。主次篇章单位的区分主要根据全局重要性做出。

目前的篇章结构理论研究最具代表性的是 RST 和 PDTB 体系，本文将 CDT 和它们进行简单的对比：1)在基本篇章单位的定义上，CDT 的基本篇章单位（子句）一定有标点作为标志，一般是小于或等于句子的单位，根据定义比较容易区分。RST 的基本篇章单位可以小到短语，PDTB 体系中连接词前面的论元记为 Arg1，后面记为 Arg2，论元可以大到多个句子，小到从句。2)在连接词的处理上，RST 没有考虑连接词，CDT 和 PDTB 体系都考虑了连接词。3)在篇章关系的处理上，CDT 和 PDTB 体系都考虑了关系类别。CDT 将关系和连接词区分开，给出一个通用的关系分类，由于有连接词标注，CDT 可以轻松的构建不同的关系体系，以便使篇章结构分析结果适用于不同任务。4)在结构树表示上，CDT 和 RST 均可构建完整的篇章结构树，PDTB 体系则没有着意构建篇章结构树，但可以根据已有关系推导出部分结构树。5)在篇章单位主次区分上，PDTB 体系不区分主次，RST 按照关系类别区分主次，CDT 按照全局重要性区分主次，属于同一种关系的两个篇章单位主次可能也不一样，例如“之所以……是因为……”和“因为……所以……”都是因果关系，但“之所以……是因为”的主要部分是原因项，“因为……所以……”的主要部分是结果项。从以上对比可知，CDT 集合了 RST 和 PDTB 的优点并结合了汉语本身的特点。

2.2 汉语篇章结构语料库

采用 CDT 的表示形式，文献[1]标注了汉语篇章结构语料(CDTB)，语料采用 XML 形式存储。目前 CDTB 共有 500 个文档，生文本选自 CTB6.0，在 CTB6.0 中句子标号从 1 到 6648。每个段落标注为一棵连接依存树，共有效标注 2342 个篇章（段落）。CDTB 共包含 10643 个子句，每棵篇章树平均 4.5 个子句。平均每个有效标注的句子包含 2 个子句。

目前 CDTB 中共有 278 个连接词，其中显式连接词有 274 个，可添加的隐式连接词有 40 个（包括部分显式连接词）。单义连接词（可表示一种关系类别）有 243 个，多义连接词（可表示多种关系类别）有 35 个。CDTB 共标注关系 7310 个，其中显式关系 1814 个（占 24.8%），隐式关系 5496 个（占 75.2%）。除标注连接词外，语料对每个关系均标注关系类型，每种关系及其出现次数在图 2 关系类型后用数字标明。

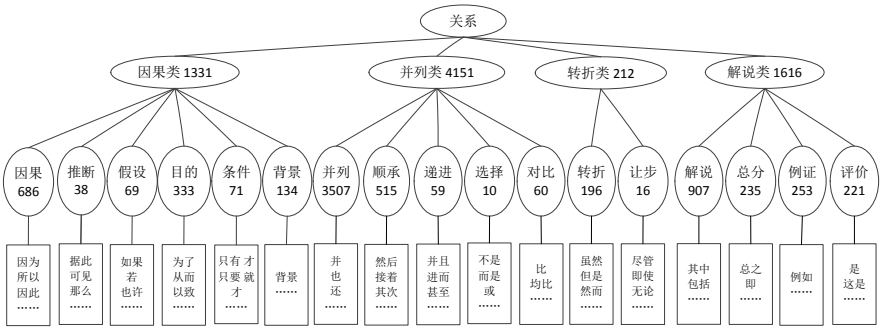

图 2 基于连接词的关系分类

四大类中，并列类有 4151 个，所占比例高达 56.8%，其中，隐式并列类有 3176 个，占所有隐式关系的 57.8%。转折类仅有 212 个实例，占 2.9%，其中，隐式转折类仅有 39 个实

例，占隐式关系的 0.7%，因为转折类仅包含转折关系和让步关系，类别较少，每种关系类别的数目也较少。

CDTB 中，层次最大为 9，层次树小于等于 4 的关系有 6958 个(95.2%)，其它层有 352 个(4.8%)：层次为 1 的关系有 2342 个(32.0%)；层次为 2 的 2372(32.5%)；层次为 3 的 1533 个(21.0%)；层次为 4 的 711 个(9.7%)。

CDTB 中，单中心的有 3555 个（48.6%），多中心的有 3755 个（51.4%）。四大类中，并列类多数情况（88.7%）下是多中心的，解说类、因果类和转折类多是单中心的。单中心情况下，中心在前的有 2108 个（59.3%），中心在后的有 1477 个(40.7%)。四大类中，因果类和转折类一般中心在后，并列类和解说类中心在前。

CDTB 一致性测试结果为：子句标注一致性为 91.7%，Kappa 值为 0.84；显式关系和隐式关系判断一致性为 94.7%，Kappa 值为 0.81；显式连接词选用一致性为 82.3%，隐式连接词添加一致性为 74.6%；主次篇章单位判定一致性 80.7%；篇章结构树标注一致性 77.4%。

3 基于语料的汉语篇章结构分析

3.1 汉语篇章结构分析框架

CDTB 语料中凡是表示子句边界的标点位置都标注有其在篇章结构树中的连接词、层次和篇章单位主次等信息。根据这些信息可以进行训练产生相应的模型，将篇章切分为子句序列，进而采用自底向上的方法自动构建篇章结构树，构建篇章结构树的同时需要判断关系类别和篇章单位主次信息。由于子句是一个连续的序列，并且只有相邻的子句才会进行组合并由特定的关系进行连接，这就大大的降低了搜索空间。

CDTB 语料中存在一个关系包含多个篇章单位的情况，如图 1 中的 def 是并列关系。由于该树并非二叉树结构，采用自底向上组合时需要将其转换成二叉结构再进行。本文通过添加虚拟节点，实验时自动将多叉树转换为向左二叉化树，树上的叶子节点为子句。本文篇章单位之间暂且只考虑四大类关系。

图 1 经过转换后如图 3 所示，并列关系的子句 def 转换成二叉树后如图 3 中虚线框所示，显式连接词和隐式连接词在图 3 中分别用它们的语义类别代替。本文汉语篇章结构分析平台的目的就是对输入文本，产生如图 3 所示的篇章结构树，该树所有的叶子节点是子句，内部节点是关系类别，篇章单位之间区分主次。

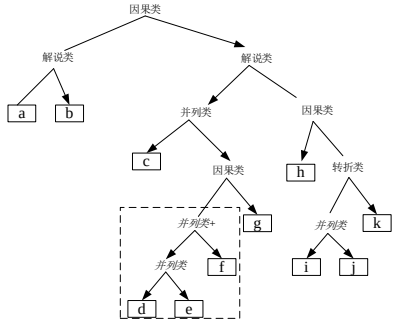

图 3 图 1 转换后的篇章结构树

为构建汉语篇章结构树，本文给出了一个汉语篇章结构分析平台（如图 4），该平台参考英语篇章结构分析平台 HILDA[21]的做法，整个平台由训练和测试两部分组成。训练部分首先对 CDTB 语料进行处理，即将 CDTB 中多叉树转换成二叉树，篇章关系转换成 4 大类，同时将句法信息和 CDTB 中的篇章标注信息对齐。训练数据处理后，分别抽取相关特征得到子句识别分类器 SegClassifier（将篇章切分为子句序列）、结构识别分类器 StrClassifier（判断相邻的子句序列之间是否存在关系）、显式连接词识别器 ExpRecClassifier（识别篇章中存在的显式连接词）和显式篇章关系分类器 ExpRelClassifier（识别显式连接词的篇章关系类型）、隐式篇章关系分类器 ImpRelClassifier（判断存在关系的篇章单位之间具体的关系类别）、

篇章主次分类器 CenClassifier（判断篇章单位的主次：在前、在后和并列）。

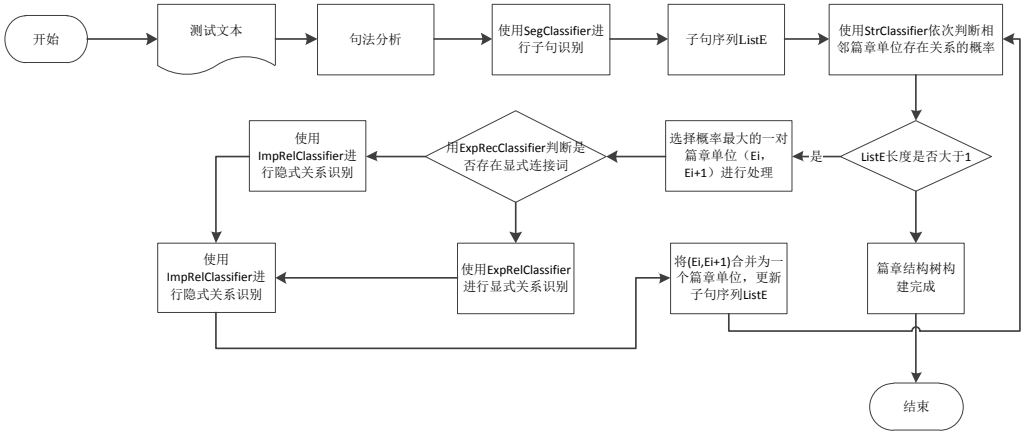

图 4 汉语篇章结构分析平台

训练完成后，测试文本进行句法分析后，首先使用 SegClassifier 进行子句识别，得到子句切分序列 ListE=[E1,E2,...]。对 ListE 中的每个元素，利用 StrClassifier 判断篇章单位相邻篇章单位之间存在关系的概率，全部计算完毕后，如果 ListE 的长度大于 1，选择概率最大的两个篇章单位(Ei，Ei+1)，判断(Ei，Ei+1)之间的篇章关系和篇章单位主次信息。判断关系前需先用 ExpRecClassifier 识别是否包含显式连接词，如包含则用显式篇章关系分类器 ExpRelClassifier 分类，否则用隐式篇章关系分类器 ImpRelClassifier 分类，篇章单位主次使用 CenClassifier 识别。将(Ei，Ei+1)合并为一个新的篇章单位 NewSpan，更新 ListE= [E0,…,Ei-1, NewSpan,Ei+2,…] 序列，以上操作循环进行，直到所有篇章单位处理完毕即得到最终的篇章结构树。本文树构建方法目前比较简单，旨在报告一个初步的汉语篇章结构分析结果。

3.2 实验方法

本汉语篇章结构分析平台主要是基于有监督的机器学习方法，因此抽取有效的特征对平台性能非常关键。由 3.1 可知，本分析平台可以拆分成多个子任务，由于每个子任务目标不一样，所用的特征也有所差别。结合前期工作，不同的任务选择不同的特征。

3.2.1 所用特征

子句识别任务主要采用文献[15]的特征，抽取每个标点的特征进行实验，具体特征及说明见文献[15]。连接词识别与分类采用文献[17]的特征和方法。篇章关系及主次识别采用文献[18]所用特征。篇章结构识别采用 Feng 等 [22]在 RSTDT 上做篇章级的结构分析时所用的特征。他们的部分特征本文在前文中已经提到，这里不再赘述。由于本文在构建篇章结构时使用的是全自动的方法，篇章关系上下文信息并不总是可用，故本文使用连接词上下文代替。本文没有采用篇章生成规则，而是在结构分析时加入以下特征：

语义相似度：本文使用哈工大《同义词词林》扩展版计算词汇的相似度，同义词词林将每个词分为大类、中类、小类，本文的方法计算子句中的词对是否在同一个中类中，如在，就抽取这个词对及其词性。

上下文信息：为当前关系中的连接词，以及前一关系及当前关系连接词的组合；为连接词及其前后词的词性；同文献[18]所述上下文特征。

依存树特征：依存树描述出各个词语之间的依存关系，即指出了词语之间在句法上的搭配关系，这种搭配关系是和语义相关联的，同文献[18]所述依存树特征。

3.2.2 实验设置

CDTB 共有 500 个文档（chtb001-chtb0657），实验统一取 450 篇文档做训练语料，50 篇文档做测试语料。为保证数据的平衡型，CDTB 中每 100 个文档取前 90 个训练，后 10 个测试。表 1 给出 CDTB 中训练集和测试集的划分及信息标注情况，表中除文档标号信息外，其它信息均为语料中统计的个数。

表1 CDTB 语料中训练和测试集划分

类别	训练集	测试集
文档数	450	50
文档标号	0001-0090, 0101-0190, 0201-0290, 0301-0325, 0400-0454, 0500-0509, 0520-0554, 0590-0596, 0600-0647	0091-0100, 0191-0200, 0291-0300, 0510-0519, 0648-0657
篇章树	2125	217
子句	9630	1013
显式关系	1657（解说类 188, 转折类 163,因果类 426,并列类 880）	157（解说类 12, 转折类 10,因果类 40,并列类 95）
隐式关系	4959（解说类 1276, 转折类 38, 因果类 786, 并列类 2859）	537（解说类 140, 转折类 1,因果类 79,并列类 317）
中心分布	单中心 3244（中心前 1901，中心在后 1343），多中心 3371	单中心 311（中心在前 207，中心在后 104），多中心 384

篇章结构分析的各项工作中，子句识别是一项基础工作。CDTB 中共有 5485 个标点位置标注有篇章信息，其中标注为子句边界的标点有 10960 个，非子句边界的标点 4525，正例占 70.8%，主要原因是正例中包含句号、分号、问号和感叹号这些句末标点，这些标点一定是子句边界。连接词识别与分类的任务是自动识别词语是否为连接词，如是连接词则对其进行关系分类。在连接词识别实验中，对语料中标注的 274 个显式连接词，我们抽取所有出现这 274 个词的例子，其中标注了连接词的为正例，没有标注的为负例。对于联合连接词（如“不但……而且”），简单起见，本文将其处理为 2 个实例，经处理后共有 226 个连接词。实验共抽取 10923 个实例，其中训练样例 10016 个，测试样例 907 个。连接词识别完成后，分别对给定连接词进行分类和自动识别连接词进行分类进行实验。在连接词分类实验中共有 2123 个实例，其中训练样例 1937 个，测试样例 186 个。在隐式篇章关系识别任务中，隐式篇章关系共有 6308 个，训练实例 5691 个，测试实例 617 个。

在系统整体平台实验上，生成训练和测试实例时进行了过滤，即一个句子内的子句不能和另一个句子内的子句或句子发生关系，如图 3 中，子句 b 和子句 c 由于不在同一个句子之中，因而不产生负例，但第 1 句（ab 的组合）和第 2 句（cdefg 的组合）产生负例。训练过程共抽取出 10554 个实例，其中存在关系（篇章单位之间存在连接）的样例有 7580 个。测试样例共有 987 个，其中存在关系的样例有 709 个。

4 实验结果及分析

4.1 基于标点的子句识别

由于句号、分号、问号和感叹号这些标点一定是子句边界，故本节实验时将其排除，剩下可能为子句边界的冒号、破折号、逗号等句内标点。句内标点子句识别整体实验结果如表 2 所示。对于标点为子句边界的情况本文记为正例，其识别 F1 值记为 F1(+)；对于标点非子句边界的情况，其识别 F1 值记为 F1(-)。具体实验结果见表 2。

表2 句内标点是否为子句边界识别结果

分类器	使用本文特征	
	标准句法树	自动句法树
	Acc. F1(+) F1(-)	Acc. F1(+) F1(-)
最大熵	93.9 95.0 92.3	91.8 93.2 89.6
决策树	74.2 81.9 55.1	73.1 81.3 51.5
贝叶斯	91.6 93.0 89.7	88.8 90.6 86.3

从表 2 可以看出，最大熵分类器在三个分类器中表现效果最好，用其作为分类器，采用标准句法树，正确率最高为 93.9%，采用自动句法树，正确率为 91.8%。对于标点为子句边界的情况，使用标准句法树和自动句法树的 F1 值分别为 95.0%和 93.2%。对于标点非子句边界的情况，使用标准句法树和自动句法树的 F1 值分别为 93.2%和 89.6%。

从表 2 可以发现，标点属于非子句边界的情况比属于子句边界的情况识别效果差，原因是占主要地位的逗号非子句边界的情况比较复杂，如子句内部主语与谓语之间的情况，子句内部动词与宾语之间的情况，这些情况判断起来比较困难，且较难找到非常有效的特征。

4.2 连接词识别与分类

显式连接词的自动识别与分类实验方法与参考文献[17]完全相同，由于实验数据划分不同，实验结果略有差别，利用最大熵分类器，给定连接词 4 大类分类结果总正确率为 95.7%，每种类别的识别结果表 3 所示。

表3　给定连接词 4 大类别识别结果

类别	准确率	召回率	F1 值
因果类	83.8	68.4	75.1
转折类	78.5	59.6	67.0
并列类	82.5	93.6	87.7
解说类	89.7	82.8	85.9

表4　自动识别连接词 4 大类识别结果

类别	准确率	召回率	F1 值
因果类	72.8	80.5	76.2
转折类	73.2	70.8	71.2
并列类	64.7	95.8	77.2
解说类	82.5	86.7	84.5

从表 3 可以发现，所有类别结果均远远好于基准系统，解说类、并列类识别效果较好，因为解说类有比较明显的连接词（如“例如”），并列类所占比例较大，识别效果也较好。转折类识别效果最差，部分原因是转折类的一些词也可以表示并列关系，例如“而”既可表示转折，又可表示并列，并列关系所占比例较大，影响了结果的判断。

通常，对连接词分类是在并不知道其是否为连接词的情况下进行，若首先使用连接词识别分类器识别实例是否为连接词，若是连接词则使用连接词分类的分类器给出类别，得到连接词分类总正确率为 89.1%，明显低于给定连接词的总正确率 95.7%。表 4 给出在连接词自动识别基础上进行连接词分类的结果，识别出连接词后对其判断类别相对容易。

4.3 隐式篇章关系识别

由实验设置可知，在训练语料中，并列类经过转换后共有 3535 个实例（占 63.0%），取结果均为概率最大的并列类为基准系统，系统正确率为 63.0%。采用最大熵分类器，4 大类隐式篇章关系识别总正确率为 66.9%，本文隐式关系识别效果好于基准系统。表 5 给出了四大类关系识别的结果，可以看出并列类识别效果最优，一方面和并列类在语料中的规模有关，另一方面，上下文特征中的共享论元模式大多数从并列类中得到，对于并列类识别有针对性。解说类识别效果次之，因果类再次之，这都和训练实例的规模有关。

表5　4 大类隐式篇章关系识别结果

类别	准确率	召回率	F1 值
因果类	37.5	19.2	24.9
并列类	72.6	85.8	78.1
解说类	54.7	45.5	49.2
转折类	--	--	--

表6　3 大类隐式篇章关系识别结果

类别	准确率	召回率	F1 值
因果类	40.6	27.7	32.4
并列类	73.7	82.3	77.3
解说类	55.9	49.1	51.8

由实验设置可知，解说实例有 1275 个，占 22.4%，因果实例有 788 个，占 13.8%。但是转折实例只有 30 个，仅占 0.6%，在测试实例中，共有关系实例 1173 个，转折类 40 个，仅占 0.7%，由此可知，转折关系在整个语料中数量较少，非常稀疏。在关系分类过程中，其会产生噪音。隐式转折关系因为数据稀疏问题没有识别出来，因此我们又考虑了去除转折类后剩余三类关系的识别情况（见表 6）。去除转折类后总正确率为 67.2%，比四大类结果提升 0.3%。

4.4 篇章单位主次识别

篇章单位主次主要是对存在关系的篇章单位，区分篇章单位之间的地位，主要有多中心、中心在前和中心在后三种类型。从表 7 可以看出，多中心识别效果最好，这和语料中多中心数据规模有关。多中心在训练语料中共有 4257 个实例，占 56.7%，在测试语料中有 485 个实例，占 60.9%。分三种类别时，总正确率为 69.0%，比测试数据偏向最大概率高 8.1%，说明本篇章单位主次识别是有用的。

表7　篇章单位主次区分识别结果

分类设置	类别	准确率	召回率	F1 值
分三类	中心在前	62.2	33.5	43.6
	中心在后	67.2	41.7	51.5
	多中心	70.4	90.8	79.3
分两类	多中心	72.7	82.1	77.1
	单中心	67.0	54.0	59.8

观察实验实例发现，中心在前的效果明显低于中心在后，而实验数据和测试数据中，中心在前的比例明显较大，说明中心在后的情况较好识别。如果考虑单中心和多中心两种类别，总正确率为 70.8%。比分成三类提高 1.8%，由于分成两类多中心所占比例下降，F1 值比分成三类下降 2.2%。

4.5 汉语篇章结构分析平台性能

汉语篇章结构分析包括子句识别、关系识别、结构树构建、主次识别等子任务。本节主要结合前面相关研究，给出汉语篇章结构分析平台的性能。平台采用 3.1 所示的分析框架，融合子句识别、连接词识别与分类、隐式关系识别、主次识别任务，采用自底向上的方法进行汉语篇章结构树的构建。下文首先分别给出结构、关系的结果，然后给出系统整体性能。

4.5.1 结构和关系识别结果与分析

结构识别主要是判断子句之间是否存在关系，关系识别主要是判断存在关系的篇章单位之间具体的关系类别，结构和关系识别均分句内、句间和综合三种情况，结果如表 8 所示。

结构识别时，句内指训练和测试数据取自句内的情况，共有 6165 个训练实例（正例 5029 个，负例 1136 个），568 个测试实例（正例 459 个，负例 109 个），句内结构识别正确率为 80.3%，识别存在关系的结构 F1 值为 87.2%。句间指训练和测试数据取自不同的句子，共有 4389 个训练实例(正例 2551 个，负例 1838 个)，419 个测试实例（正例 250 个，负例 169 个），结构识别正确率为 85.4%。综合指将以上句内句间两种情况综合训练和测试，正确率为 81.1%。从表 8 可以看出，有关系的结构识别结果明显高于无关系的结构，综合识别正确率介于之间。

表8　结构识别结果

组合	有关系				无关系		
	正确率	准确率	召回率	F1 值	准确率	召回率	F1 值
句内	80.3	80.0	95.7	87.2	81.3	43.7	56.9
句间	85.4	85.6	96.4	90.7	84.2	54.1	65.9
综合	81.1	81.2	95.7	87.9	80.0	43.4	56.3

比较表 8 句内和句间的结果我们发现，句间结构识别的正确率高于句内，正确率分别为 85.4%和 80.1%。这个结果和英语的实验结果不同，文献[22]中，英语的句内结构的识别结果均要好于句外。一方面是英语的句子相对较短，结构清晰，汉语句子相对较长，并且复句所占比例较多，复句内部结构复杂；另一方面，大部分英语句子是主从结构，主语和从句非常清晰，但汉语省略较多，句子结构也比较复杂。如例 1 中的第 2 句共有 107 个字，5 个子句，3 个关系，句内结构有三层，所以构建句 2 内部的结构树也是相当困难的任务。

关系识别是对存在关系(包含显式和隐式)的篇章单位之间的具体关系类别进行 4 大类。实验结果见表 9。句内关系识别效果最好，句间效果最差。并列关系识别效果最好，句内并列关系 F1 值可以达到 87%，因果类次之，转折类识别效果最差，这和转折类数据较少有关。对比表 9 和表 5 可知，同时考虑隐式关系和显式关系的识别效果比只考虑隐式关系要好。

综合实验的结果显示，结构和关系的识别结果均低于句内，高于句间。这说明对于结构和关系的识别可分为句内和句间分别处理，采用针对性的方法提高句间关系识别的性能。

表9 4 大类关系识别结果

组合	正确率	并列类			解说类			因果类			转折类		
		准确率	召回率	F1	准确率	召回率	F1	准确率	召回率	F1	准确率	召回率	F1
句内	78.4	79.7	95.8	87.0	50.0	21.3	29.8	82.3	45.2	58.3	100.0	25.0	40.0
句间	69.6	72.7	91.6	81.1	50.0	44.0	46.8	73.9	34.7	47.2	100.0	5.2	4.7
综合	76.2	77.5	96.0	85.8	54.3	31.9	40.2	84.2	43.2	57.1	67.0	8.7	15.4

表 10 给出篇章结构树构建的结果。本文认为当且仅当系统判断有关系的实例和所标注的语料中的实例完全一致时，这个“结构”是正确的，即自动识别结构的左右子句和标注篇章中结构的左右子句内容完全一致。“结构+关系”是正确识别结构的基础上，判断此结构对应的自动识别的四大类关系和标注的四大类关系是否一致。“结构+主次”是在结构一致的情况下，判断篇章单位主次是否一致，篇章单位主次分为多中心、左中心和右中心三种类别。“结构+关系+主次”是在结构一致的基础上，关系和中心判断也一致。表 10 分四种情况实验，标准子句指语料中手工标注的子句，自动子句是基于句内标点自动识别的子句。

表10 整个系统的性能

组合	结构			结构+关系			结构+主次			结构+关系+主次		
	准确率	召回率	F1	准确率	召回率	F1	准确率	召回率	F1	准确率	召回率	F1
1 标准子句和标准句法树	54.9	56.3	55.6	33.8	34.3	34.5	25.8	26.5	26.2	24.0	24.5	24.2
2 标准子句和自动句法树	51.7	53.0	52.3	33.4	34.3	33.8	22.6	25.4	23.9	22.9	23.4	23.2
3 自动子句和标准句法树	46.0	51.5	48.6	27.5	30.7	29.0	21.8	24.1	23.1	19.9	22.2	21.0
4 自动子句和自动句法树	44.0	49.1	46.4	27.3	30.5	28.8	21.5	24.8	23.1	19.0	21.2	20.0

表 10 中分别给出对应情况下的准确率、召回率和 F1 值。第 1 行给出了采用目前特征可以取得的最好结果，采用语料中标注的子句和标准句法信息，“结构”识别 F1 值为 55.6%，“结构+关系”识别的 F1 值为 34.5%，“结构+主次”F1 值为 26.2%，“结构+关系+主次”F1 值为 24.2%。第 4 行给出完全自动方法所得到的结果，“结构”识别 F1 值为 46.4%，“结构+关系”识别的 F1 值为 28.8%，“结构+主次”F1 值为 23.1%，“结构+关系+主次”F1 值为 20.0%，F1 值分别比最好的情况低 9.2%、5.7%、3.1%和 4.2%。整体来说，结构识别效果最好，其次是“结构+关系”结果，再次是“结构+主次”，效果最差也是最难的是得到图 2 所示的“结构+关系+主次”篇章结构树。

对比第 2 和第 4 种情况，可以发现，2 和 4 都采用的是自动句法树，不同的是 2 采用标准子句而 4 采用自动识别的子句，最后的结果 4 比 2 的 F1 值分别低了 5.9%、5.0%、0.8%和 3.2%，这个结果也充分说明子句识别是篇章结构树构建的基础。虽然本文的子句识别正确率基本达到了 90%，但开始的错误会逐级向下放大传递。

对比第 1 和第 2 种情况，虽然都使用标准子句，但第 1 种情况使用标准句法树，F1 值分别比第 1 种情况高 3.3%、0.7%、2.3%和 1.0%，表明句法树对结果是有影响的。

英语中基于 RST 语料的篇章结构分析目标也是构建篇章结构树，和本文非常相似，目前报告的最好结果是：结构分析的 F1 值为 69.8%[22]，关系分类（18 类）的 F1 值为 65.1%[21]，“结构+关系”F1 值为 47.8%[22]，“结构+主次”59.1%[22]，“结构+关系+主次”F1 值为 47.3%[22]。由于本平台是首次在 CDTB 上的分析尝试，没有相关的实验可以对比，但能看出汉语篇章结构分析研究还很初步，仍然有很多工作要做。

5 总结

本文主要进行基于连接依存树的汉语篇章结构分析平台构建。该平台包括子句识别、连接词识别与分类、篇章关系识别和篇章单位主次识别几个子任务。平台使用自底向上的方法进行篇章结构树构建，输出是简化的连接依存树。实验结构表明基于连接依存树的汉语篇章

结构表示体系是合理的，基于此表示体系构建的汉语篇章结构语料 CDTB 是可用的。这是汉语篇章结构系统化分析的首次尝试，为后续研究提供了基础平台。

本文汉语篇章结构分析平台整体效果还不尽如人意，下一步准备参考句法分析的做法采用全局优化的进行篇章结构分析，从而提高篇章结构平台整体性能。

参考文献

[1]Li YC, Feng WH,Sung J et al.Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure[C]. In: Proc.of EMNLP 2014:2105-2114.
[2]Carlson L, Marcu D, Okurowski ME. Building a discourse-tagged corpus in the framework of rhetorical structure theory[M]. Berlin: Springer Netherlands,2003:85-112.
[3]PDTB-Group. The Penn Discourse Treebank 2.0 annotation manual[R]. Technical Report IRCS-08-01, Institute for Research in Cognitive Science, University of Pennsylvania, 2008.
[4]乐明. 汉语篇章修辞结构的标注研究[J]. 中文信息学报,2008,22(4):19-23.
[5]陈莉萍. 英汉语篇结构标注理论与实践[J]. 上海:上海外国语大学,2006.
[6]Xue NW. Annotating the Discourse Connectives in the Chinese Treebank[C]. In: Proc. of the ACL Workshop on Frontiers in Corpus An-notation, 2005. 84-91.
[7]Zhou YP, Xue NW. PDTB-style discourse annotation of Chinese text[C]. In: Proc. of the 50th Annual Meeting of the ACL, Association for Computational Linguistics, 2012. 69-77.
[8] Huang HH, Chen HH. Contingency and comparison relation labeling and structure prediction in Chinese sentences[C]. In: Proc. of the 13th Annual Meeting of the SIGDIAL,2012. 261-269.
[9]Zhou, LJ, Li BY, Wei ZY, et al. The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank[C]. In: Proc. of the ICLRE,2014: 942-949.
[10]张牧宇,秦兵,刘挺. 中文篇章级关系体系及类型标注[J]. 中文信息学报, 2014,28(2):28-36.
[11]邢福义.汉语复句研究[M]. 北京:商务印书馆, 2001.
[12]周强.汉语句法树库标注体系[M]. 中文信息学报, 2004, 18(4):1-8.
[13]Xue NW, Yang YQ. Chinese sentence segmentation as comma classification[C]. In: Proc. of the 49th Annual Meeting of the ACL, Association for Computational Linguistics, 2011. 631-635.
[14]Yang YQ, Xue NW. Chinese comma disambiguation for discourse analysis[C]. In: Proc. of the 50th Annual Meeting of the ACL, Association for Computational Linguistics, 2012:786-794.
[15]李艳翠,冯文贺,周国栋,等. 基于逗号的汉语子句识别研究[J]. 北京大学学报(自然科学版).2013, 49(1):7-14.
[16]李艳翠,孙静,周国栋,等. 基于清华汉语树库的复句关系词识别与分类研究[J]. 北京大学学报(自然科学版).2014, 50(1):118-124.
[17]李艳翠,孙静,周国栋. 汉语篇章连接词识别与分类[J]. 北京大学学报(自然科学版),2015,51(1):7-14.
[18]孙静,李艳翠,周国栋,等. 汉语隐式篇章关系识别[J]. 北京大学学报(自然科学版), 2014,50(1):112-117.
[19]张牧宇,宋原,秦兵,等. 中文篇章级句间语义关系识别[J]. 中文信息学报, 2013, 27(6):51-57.
[20]涂眉,周玉,宗成庆. 基于最大熵的汉语篇章结构自动分析方法[J]. 北京大学学报(自然科学版), 2014,50(1):125-132.
[21]Hernault H., Helmut P., David A. D., et al. HILDA: A discourse parser using support vector machine classification[J]. Dialogue and Discourse, 2010:1(3):1–33.
[22]Feng V. W. and Hirst G. Text-level discourse parsing with rich linguistic features[C]. In Proc. Of ACL, 2012: 60-68.

作者简介：

李艳翠(1982－)，女，讲师，博士研究生，主要研究领域为自然语言处理。
Email: yancuili@gmail.com

孙静(1986－)，女，博士研究生，主要研究领域为自然语言处理。
Email:sj44581@163.com

冯文贺(1976－)，男，讲师，博士，主要研究领域为计算语言学。
Email：Wenhefeng@gmail.com

周国栋(1967－)，男，教授，博士，博士生导师，主要研究领域为自然语言处理、多语言跨文本信息抽取。
通讯作者, Email:
gdzhou@suda.edu.cn