



博士学位论文



论文题目 汉语篇章结构表示体系及资源构建研究

研究生姓名 李艳翠

指导教师姓名 周国栋（教授）

专业名称 计算机应用技术

研究方向 自然语言处理

论文提交日期 2015 年 9 月

苏州大学学位论文独创性声明

本人郑重声明：所提交的学位论文是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含其他个人或集体已经发表或撰写过的研究成果，也不含为获得苏州大学或其它教育机构的学位证书而使用过的材料。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人承担本声明的法律责任。

研究生签名：_____日 期：_____

苏州大学学位论文使用授权声明

本人完全了解苏州大学关于收集、保存和使用学位论文的规定，即：学位论文著作权归属苏州大学。本学位论文电子文档的内容和纸质论文的内容相一致。苏州大学有权向国家图书馆、中国社科院文献信息情报中心、中国科学技术信息研究所（含万方数据电子出版社）、中国学术期刊（光盘版）电子杂志社送交本学位论文的复印件和电子文档，允许论文被查阅和借阅，可以采用影印、缩印或其他复制手段保存和汇编学位论文，可以将学位论文的全部或部分内容编入有关数据库进行检索。

涉密论文 ☐
本学位论文属 在_____年_____月解密后适用本规定。
非涉密论文 ☐

论文作者签名：_____日 期：_____

导 师 签 名：_____日 期：_____

汉语篇章结构表示体系及资源构建研究

中文摘要

篇章指由一系列连续的子句、句子或语段构成的语言整体单位,每个篇章不仅具有内部连贯性,而且篇章中的各级语言单位是描述同一问题或同一种情境的一个相对完整的语言整体。在一个篇章中,子句、句子或语段间具有一定的**层次结构和语义关系**,篇章结构分析旨在分析出其中的层次结构和语义关系,篇章结构分析结果对于**提高自动文摘、知识抽取、自动问答等相关应用系统的性能均有重要作用**。随着词法、句法分析技术的成熟,篇章结构分析成为制约自然语言处理的关键基础问题。

目前篇章结构分析研究主要面向英语,而面向汉语的研究相对落后。主要表现在:1) 适用于汉语的篇章结构分析理论还不完善;2) 符合汉语特点的大规模汉语篇章结构语料资源匮乏;3) 由于汉英语言上的差异性,适用于英语的篇章结构分析方法不能直接应用在汉语篇章结构分析研究中。

本文重点研究了**汉语篇章结构的理论表示体系**。借鉴英语修辞结构理论和宾州篇章树库体系的优点,参考汉语复句和句群的研究成果,结合汉语本身特点,本文提出一种**基于连接依存树的汉语篇章结构表示体系**,并根据汉语特点定义了其中的关键元素:**子句(基本篇章单位)、连接词、篇章结构关系、篇章单位主次**。连接依存树的主要特征是叶子节点为子句,内部节点为连接词,连接词通过其层级地位(管辖范围)表示篇章结构的层次,通过其语义(具体与抽象)表示篇章关系,连接词所连接的篇章单位根据篇章整体意图区分主次。与修辞结构理论、宾州篇章树库体系的理论对比表明,本文所提基于连接依存树的汉语篇章结构表示体系在理论上具有一定的优越性,并且符合汉语特点。基于连接依存树的汉语篇章结构表示体系是进一步开展篇章结构语料库构建的理论基础。

在此基础上,进行了汉语篇章结构语料库构建研究。基于连接依存树表示体系,本文采用自顶向下的标注策略和人机结合的语料库标注方式,构建了汉语篇章结构语料库(Chinese Discourse Treebank, **CDTB**)。CDTB 目前包含来自 Chinese Treebank 的 500 个文档,本文对其进行了分析并展示了语料库的标注情况。标注一致性测试表明语料质量较好,统计数据表明所标资源达到一定规模。CDTB 可以为汉语篇章结构分析研究提供资源支持。

最后,本文构建了基于连接依存树的汉语篇章结构分析平台。该平台包括子句识别、篇章结构树构建、篇章关系识别、篇章单位主次识别任务,实验结果验证了本文提出的基于连接依存树表示体系的合理性和所标注的 CDTB 语料库的可用性。

目前,汉语篇章结构分析研究尚处于起步阶段,本文研究亦属探索性工作,上述工作在理论研究、资源建设、计算分析上对汉语篇章结构分析均有不同程度的创新,对该领域的相关研究具有重要参考价值。

关键词: 篇章结构分析, 连接依存树, 语料库, 子句, 篇章关系

作 者: 李艳翠

指导教师: 周国栋

Research of Chinese Discourse Structure Representation and Resource Construction

Abstract

It is well-known that interpretation of a discourse requires understanding of its rhetorical relation hierarchy since discourse units rarely exist in isolation. Research in discourse parsing is aim to reveal such relations hierarchy in discourse, which is helpful for many downstream applications, including summarization, information retrieval and question answering, etc. Due to the maturity of the lexical and syntactic parsing technologies, discourse parsing has been attracted more and more attentions in recent years.

In comparison with English, however, there are rare studies on Chinese discourse parsing. The main reasons include: 1) There is no complete theory for Chinese discourse parsing; 2) The corpus of Chinese discourse which matches the characteristic of Chinese discourse structure is deficient; 3) Due to the complexity of Chinese discourse structure, the existing methods can not be directly applied on Chinese discourse parsing.

This dissertation focuses on Chinese discourse structure representation theory. We first take value of various theories and representation scheme on the tree structure and nuclearity of Rhetorical Structure Theory (RST), relation and discourse structure of Chinese compound sentence and the sentence-group theory, the connective treatment of Penn Discourse Tree Bank (PDTB). Then we propose an effective discourse representation scheme for Chinese, called Connective-driven Dependency Tree (CDT), and give the definition of clause, connective, relation and nuclearity etc. In CDT, clauses are regarded as leaf nodes while connectives are regarded as non-leaf nodes. In particular, connectives directly represent the hierarchy of the tree structure and the rhetorical relation of a discourse, while the nucleus of discourse units is globally determined according to the dependency theory. Compared with RST and PDTB, CDT representation has certain advantages to meet the special characteristics of Chinese discourse structure. Therefore, CDT representation is the basis of discourse corpus construction.

Guided by the CDT scheme, we construct Chinese discourse structure corpus, i.e., Chinese Discourse Treebank (CDTB). This is done by manually annotating 500 documents in Chinese Treebank (CTB). We use a top-down segmentation strategy and incorporate with both manual and automatic annotation approach to keep consistency with Chinese natives' cognitive habits. Then we show the statistics and analysis of CDTB in details. The consistency test shows that the CDTB quality is good. And the statistical data shows that the CDTB reaches an available size. Therefore, CDTB can provide resource for the task of Chinese discourse parsing.

Finally, this dissertation presents a Chinese discourse parser based on our CDTB corpus. The input of the platform is raw text while the output is a discourse tree, including the clause, the discourse hierarchy structure, discourse relations (4 classes) and discourse nuclearity. Experimental results show the appropriateness of the CDT scheme of Chinese discourse analysis and the effective of our CDTB corpus.

The current studies on Chinese discourse structure is still in primary stage. The research of the Chinese discourse parsing on theory, resource, computing has great innovation in Chinese discourse parsing. The research work exhibits a great reference value to the future research in Chinese discourse parsing.

Keywords: Discourse parsing, Corpus, Clause, Connective, Discourse relation

Written by Li Yancui

Supervised by Zhou Guodong

目录

第 1 章 绪论.....	1
1.1 研究背景和意义.....	1
1.2 国内外研究现状.....	3
1.2.1 英语篇章结构分析的理论研究.....	4
1.2.1.1 浅层的衔接关系.....	4
1.2.1.2 Hobbs 模型.....	4
1.2.1.3 修辞结构理论.....	5
1.2.1.4 宾州篇章树库体系.....	7
1.2.1.5 其它相关理论.....	9
1.2.2 英语篇章结构分析的资源建设.....	9
1.2.2.1 修辞结构理论篇章树库.....	9
1.2.2.2 宾州篇章树库.....	11
1.2.3 英语篇章结构分析的计算模型.....	13
1.2.3.1 基于 RSTDT 的研究.....	14
1.2.3.2 基于 PDTB 的研究.....	16
1.2.3.3 结合 RSTDT 和 PDTB 的研究.....	18
1.2.4 汉语篇章结构分析研究现状及存在问题.....	18
1.2.4.1 汉语篇章结构分析研究现状.....	18
1.2.4.2 存在问题及研究趋势.....	21
1.3 本文的研究内容.....	21
1.4 本文的组织结构.....	22
第 2 章 基于连接依存树的汉语篇章结构表示体系.....	23
2.1 引言.....	23
2.1.1 已有篇章结构理论体系分析.....	23
2.1.2 汉语篇章结构的特点.....	24
2.1.3 连接依存树.....	24
2.2 叶子节点—子句.....	26

2.2.1 子句的定义.....	26
2.2.2 子句的判定.....	28
2.2.2.1 子句是单句.....	28
2.2.2.2 子句是复句中的分句.....	28
2.2.2.3 标点与子句判定.....	29
2.2.2.4 一些特殊情况.....	35
2.3 内部节点—连接词.....	36
2.3.1 连接词的特点.....	37
2.3.1.1 连接词的形式.....	37
2.3.1.2 连接词的分布.....	38
2.3.1.3 连接词的词性.....	39
2.3.1.4 连接词的句法特性.....	39
2.3.1.5 连接词的逻辑语义关系.....	39
2.3.1.6 其它连接词.....	40
2.3.2 隐式连接词的添加.....	40
2.3.2.1 添加连接词的依据.....	41
2.3.2.2 连接词添加的位置.....	43
2.3.2.3 其它情况.....	43
2.3.3 显式连接词的删除.....	45
2.4 篇章结构关系.....	46
2.4.1 篇章结构层次化及判定.....	46
2.4.2 篇章关系类别及判定.....	48
2.4.2.1 篇章关系类别.....	49
2.4.2.2 篇章关系的判定.....	53
2.5 篇章单位主次.....	55
2.5.1 篇章单位主次区分.....	55
2.5.2 篇章单位主次判定.....	57
2.5.2.1 主次判定的依据.....	57
2.5.2.2 主次判定的方法.....	60
2.5.2.3 主次判定的难点.....	61

2.6 与相关理论的比较	62
2.7 本章小结	63
第 3 章 基于连接依存树表示体系的 CDTB 语料库构建	65
3.1 引言	65
3.2 自顶向下的 CDTB 标注策略	65
3.3 人机结合的 CDTB 标注方法	66
3.3.1 标注流程设计	67
3.3.2 语料标注	67
3.3.3 语料格式	69
3.3.4 语料校对	70
3.4 CDTB 标注一致性测试	70
3.5 CDTB 标注信息统计与分析	72
3.5.1 连接词统计与分析	73
3.5.2 篇章关系统计与分析	74
3.5.3 篇章结构统计与分析	76
3.5.4 篇章单位主次统计与分析	78
3.6 本章小结	80
第 4 章 基于 CDTB 的汉语篇章结构分析	81
4.1 引言	81
4.2 汉语篇章结构分析框架	83
4.3 实验方法	85
4.3.1 所用特征	85
4.3.1.1 子句识别	86
4.3.1.2 连接词识别与分类	87
4.3.1.3 篇章关系及主次识别	88
4.3.1.4 篇章结构识别	90
4.3.2 实验设置	90
4.4 实验结果及分析	92
4.4.1 基于标点的子句识别	92

4.4.2 连接词识别与分类.....	95
4.4.2.1 连接词识别.....	95
4.4.2.2 连接词分类.....	96
4.4.3 隐式篇章关系识别.....	97
4.4.4 篇章单位主次识别.....	99
4.4.5 基于连接依存树的汉语篇章结构分析平台性能.....	99
4.4.5.1 结构和关系识别结果与分析.....	100
4.4.5.2 篇章结构树构建结果与分析.....	101
4.5 本章小结	103
第 5 章 总结与展望.....	104
5.1 总结	104
5.2 展望	105
参考文献.....	106
作者在攻读博士学位期间完成的论文及科研工作.....	114
附 录.....	116
致 谢.....	126

第1章 绪论

1.1 研究背景和意义

自然语言的单位由小到大可以分为词、短语、句子和段落，最后形成篇章（Discourse）^[1]，篇章有时也称语篇或话语。在篇章结构分析中，篇章指一系列连续的句子、句子或语段构成的语言整体单位。篇章不是语言成分的无序堆砌，每个篇章不仅具有内部连贯性，而且篇章中的各级单位都是描述同一个问题或同一种情境的一个相对完整的语言整体。也就是说，虽然在形式上篇章由句子序列构成，但句子序列并不一定能构成篇章。如例 1.1 中，虽然单个句子都是正确的，句法上单个句子也是完整的，但是这些句子顺次组合在一起并没有形成篇章。因为例 1.1 中的句子在意义上不关联，没有表达明确的主题，所以无法形成一个整体。与此相比，例 1.2 中，虽然有的分句并不完整（用【】表示缺省成份，也称为零指代），但分句前后关联，主题清晰，整体上围绕“张三”展开，构成了一个篇章。

一个篇章中，子句、句子或语段间具有一定的层次结构和语义关系，只有分析出其中的层次结构及语义关系，才能对篇章有一个总体把握。篇章结构分析目的是分析出篇章的层次结构及语义关系，它是自然语言处理的一个核心问题，也是近几年的一个研究热点和难点。图 1.1 给出了一个篇章（例 1.2）的层次化结构实例。

例1.1 比尔来自于美国。今天交通非常拥挤。长江贯穿中国的多个省市。因此，自然语言处理是计算机科学与语言学的融合。

例1.2 a 张三才 30 出头，||b【】<而且>既没有什么学历，|||c【】又没有什么新的工作经验。|d 但是【】不论干什么，|||e 他都非常认真，||f 所以，领导总是把一些重要的任务交给他。

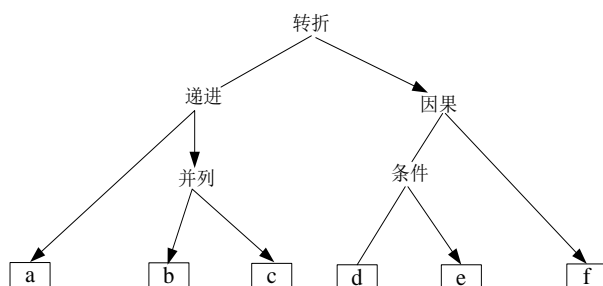


图1.1 例 1.2 的篇章层次化结构实例

例 1.2 中的字母表示基本篇章单位，“|”的数量代表篇章单位间的组合层次。图

1.1 中箭头所指向的内容为**主要篇章单位（中心）**。其中 b 和 c 之间是并列关系，b 和 c 都是中心，bc 组合和 a 是递进关系，abc 组合和 def 组合是转折关系，def 组合为中心。各基本篇章单位组合后形成高级篇章单位，进而通过再组合形成更高级篇章单位，如此层层组合，最终可以形成一棵**篇章结构树**¹。各层篇章单位赖以组合的原因在于其间存在一些为数不多的、反复出现的结构关系（如并列、递进等），这些结构关系有时有连接成分作为形式标记（如例 1.2 中的“既……又……”），有时则完全隐含（如例 1.2 中的“而且”）。

篇章结构分析结果有利于理解篇章，同时对自然语言处理应用也非常重要，例如：对于自动问答，利用如图 1.1 中的**因果关系**能够抽取出问答系统的问题和答案，“领导总是把一些重要的任务交给他”的原因是“不论干什么，他都非常认真”；对于自动摘要，根据篇章结构分析的结果，箭头所指内容为中心，def 组合比 abc 组合重要，若抽取部分篇章内容，根据重要程度不难得到 def 组合，若只抽取一个基本篇章单位，则根据中心指向可以得到 f。

自然语言处理研究尽管已经经历了几十年的发展历程，但一般重点通常聚焦在句法和词法层面，对篇章内在规律的研究相对较少，缺乏对篇章结构进行分析的理论和计算方法，从而严重制约了基于篇章的相关应用。可以说，目前的篇章结构分析研究，无论在理论上还是在实践上，都处于起始阶段。尤其对于汉语的篇章结构分析研究，还没有建立起与之相适应的理论体系和计算方法，基础资源十分匮乏。

鉴于此，本文的**汉语篇章结构分析**研究具有非常重要的理论意义和应用价值。**理论上**，本文提出**基于连接依存树的汉语篇章结构表示体系**，有助于推动汉语篇章结构分析的理论体系研究，丰富和发展语言学理论；**资源上**，**本文构建基于连接依存树表示体系的汉语篇章结构标注语料库**，可为汉语篇章结构分析及其它相关研究提供基础数据；**计算分析上**，**本文给出了汉语篇章结构分析平台**，并给出了相应的子句识别、连接词识别、篇章关系识别、篇章主次识别和篇章结构树构建的基本方法，为后续汉语篇章结构分析研究提供参考；**应用上**，本文汉语篇章结构分析结果将有助于提高自动文摘、知识抽取、问答系统等相关应用的性能，从而推进自然语言处理基础研究和应用研究的开展。

¹篇章单位之间可能会出现交叉的情况，不过这种情况不多，可以参照依存句法分析的方法予以合理解决

1.2 国内外研究现状

篇章分析是自然语言处理的核心问题。早在 20 世纪 70 年代,语言学家和认知科学家就对这个问题开展了研究。Schank 和 Abelson^[2]首先提出了著名的概念依存理论 (Concept Dependency),并在此基础上提出了脚本 (Script) 方法,对特定的“故事”进行理解。目前备受关注的信息抽取研究就采取类似的思想,只是简化了抽取的内容。但是脚本方法过于依赖领域,当场景发生变化时,就需要构建新的脚本。而很多篇章很难用特定的场景来描述,这大大限制了脚本方法的推广使用。

篇章分析需要用更加通用和开放的表达形式来处理。这就需要充分挖掘篇章的一般知识,明确篇章的基本特征。Beaugrande 和 Dressler^[3]认为篇章具有衔接性 (cohesion)、连贯性 (coherence)、意图性 (intentionality)、可接受性 (acceptability)、信息性 (informativity)、情景性 (situationality) 和跨篇章性 (intertextuality) 七个基本特征。其中,衔接性、连贯性、意图性和信息性四个基本特征对自然语言处理产生了深远的影响^[4-12]。衔接和连贯常常以表层形式体现,为篇章分析提供了“形式标记”。与此相比,信息性和意图性属于篇章语义层面的特征,隐藏在篇章更深的层次上,通常可以融合在连贯性中考虑。信息性强调文本的内容,是作者期望向读者传达的(新)信息;而意图性强调作者的写作意图,是作者期望通过传达信息对读者形成的某种影响。

无论是西方语言还是汉语,篇章的衔接性和连贯性都是最需要关注的两个问题,是篇章的两个最基本特征。De Beaugrande 和 Dressler^[13]认为,衔接指的是形式联系,是表层篇章成分之间有顺序的相互联系的方法,主要表现为整个篇章范围内词汇(或短语)之间的关联,其中表层篇章其实就是我们所见到和听到的实际词语。与衔接不同,连贯指的是功能联系,主要通过句子(或句群)之间的语义关联来表示篇章的关联。从这个意义上看,连贯是一种内部连接,正是有了内部连接,才使得篇章具有整体性。本质上,衔接性和连贯性分别从内容和表达两个方面保证了篇章的正确性和可理解性,二者相互依赖,相互补充。

由于英语篇章结构分析研究起步较早,研究成果较多,本文所述主要是针对语法结构较强的书面语。下面首先从理论研究、资源建设、计算模型三方面分述英语篇章结构分析的研究现状,然后介绍汉语篇章结构分析的相关研究情况。

1.2.1 英语篇章结构分析的理论研究

英语篇章结构分析的理论研究比较多,本文总结并给出比较重要且有实际应用的相关理论,从而为汉语篇章结构分析提供参考。相关理论主要包括 Hobbs 模型^[6-7]、修辞结构理论^[8-10]和宾州篇章树库体系^[14-15]等。

1.2.1.1 浅层的衔接关系

在语言学研究中, Halliday 等^[16]最早将衔接的各种修辞手段作为一种专门的语言现象进行系统的分析和详尽的研究。他们合著的《Cohesion in English》标志着衔接理论的创立,是语言学理论创建的有益探索。衔接是一个语义概念,它是指篇章中语言成分之间的语义联系,或者说是篇章中一个成分与另一个可以与之相互解释的成分之间的关系。当篇章中一个成分的含义依赖于另一个成分的解释时,便产生了衔接关系。作者提出了衔接手段的分类:语法衔接(grammatical cohesion),包括照应(reference)、替代(substitution)、省略(ellipsis)和连接(conjunction);词汇衔接(lexical cohesion),包括重复(repetition)、同义/反义(synonymy/antonymy)、上下义/局部-整体关系(hyponymy/meronymy)和搭配(collocation)。即连接是运用连接成分体现篇章不同成分之间具有何种逻辑关系的手段。作者将句子作为篇章连接的基本单位,初步将连接成分划分为四种类型,即加合(additive)、转折(contrastive)、因果(causal)和时间(temporal)。Martin^[17]研究连接关系,他认为连接关系存在于子句(不仅仅是句子)之间,并给出了隐式关系的概念,即子句可以插入连接词。隐式关系在之后的宾州篇章树库标注中被广泛采用。Grimes^[18]考虑非词汇化的命题关系,除了给出更详细的连贯关系类别,他还介绍了被修辞结构理论采用的并列(paratactic)关系和主从(hypotactic)关系,其中,并列关系即关系的论元同等重要,主从关系即关系的论元有主次之分。

1.2.1.2 Hobbs 模型

Hobbs 模型^[6-7]提出,篇章结构由篇章单位(Discourse Unit)和连贯关系(Coherence Relation)构成。其中篇章单位可以小到子句,大到篇章本身;连贯关系是两个篇章单位之间的语义关联性,包括12种关系类型,如原因(Cause)、背景(Background)、细化(Elaboration)等。设S0和S1为两个相关的句子,部分篇章关系定义如下:结果关系(Result):推测S0所声明的状态或事件(可能)导致S1所声明的状态或事件;解释关系(Explanation):推测S1所声明的状态或事件(可能)导致S0所声明

的状态或事件；**并列关系**（Parallel）：推测 S0 所声明的 $P(a_1, a_2, \dots)$ 与 S1 所声明的 $P(b_1, b_2, \dots)$ 是类似的；**细化关系**（Elaboration）：推测 S1 和 S0 所声明的是同一命题 P；**时机关系**（Occasion）：推测由 S0 所声明的状态到 S1 最终状态的变化，或者由 S1 所声明的状态到 S0 的最初状态的变化。

Hobbs 提出的连贯关系及语义表示对其他研究者有较大的影响，如宾州篇章树库体系^[15]就借鉴了 Hobbs 模型的成果。

1.2.1.3 修辞结构理论

修辞结构理论（Rhetorical Structure Theory, RST）是美国学者 Mann 和 Thompson^[8-10]在系统功能理论的框架下创立的篇章生成和分析的理论。**RST 旨在“描述那些使篇章成为人类交际有效的和能理解的工具的功能和结构”。**

RST 认为篇章各小句不是杂乱无章的堆放在一起的，而是存在各种各样的修辞关系，它共定义了 **20 多种修辞关系**。每个修辞关系的定义包括**限制条件（Constraints）**和**效果（Effect）**两个部分：限制条件包括核心结构段限制条件、辅助结构段限制条件及这两种结构段联合限制条件；效果指对作者使用某一关系所达到的效果及效果轨迹的说明。每个修辞关系可以连接两个或多个篇章单位。通常修辞关系连接的单位存在主次之分，其中表示主要信息的单位称作“**核（Nucleus）**”，表达次要信息的单位称作“**卫星（Satellite）**”，这类关系也称为“单核”修辞关系。此外，也有一些修辞关系连接的单位无主次之分，如对比关系（Contrast）和序列关系（Sequence），这类关系称为“多核”关系。图 1.2 给出了 RST 中证据关系定义，例 1.3 给出了一个证据关系实例。

关系名称：证据关系 核限制条件：无 卫星限制条件：无 核-卫星限制条件： 对卫星的理解导致对核相信程度的增加 效果：读者对核的可信程度增加 效果轨迹：核心

图1.2 RST 中证据关系定义

- 例1.3 e1 The next music day is scheduled for July 21(Saturday),moon-midnight.
 e2 I'll post more details later,
 e3 but this is a good time to reserve the place on your calendar.

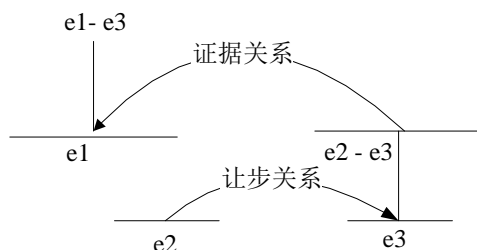


图1.3 例 1.3 的 RST 结构图示

图 1.3 给出了例 1.3 的 RST 结构图示，最上面的 e1-e3 表示结构段范围，横线下面的 e1、e2 和 e3 代表结构段，**弧线代表结构段之间的关系**，在核心-卫星关系中，箭头指向**核心**结构段，在非核关系中，弧线不带箭头。

修辞关系具有相对开放性的特点，这些关系是 RST 的核心，也是篇章连贯的重要标志。图 1.4 给出了 RST 中定义的篇章关系类型，**单核关系中证明关系、背景关系、条件关系等核心段在后面**，**详述关系、目的关系、重述关系等核心段在前**，对比关系和序列关系等是多核心关系。

Circumstance(环境关系)	Antithesis and concession
Solutionhood(解答关系)	Antithesis(对照关系)
Elaboration(阐述关系)	Concession(让步关系)
Background(背景关系)	Condition and Otherwise
Enablement and Motivation	Condition(条件关系)
Enablement(使能关系)	Otherwise(析取关系)
Motivation(动机关系)	Interpretation and Evaluation
Evidence and Justify	Interpretation(解释关系)
Evidence(证据关系)	Evaluation(评价关系)
Justify(证明关系)	Restatement and Summary
Relation of Cause	Restatement(重述关系)
Volitional Cause(意愿性原因关系)	Summary(综述关系)
Non-Volitional Cause(非意愿性原因关系)	Other Relation
Volitional Result(意愿性结果关系)	Sequence(序列关系)
Non-Volitional Result(非意愿性结果关系)	Contrast(对比关系)
Purpose(目的关系)	

图1.4 RST 中的修辞关系

在 RST 中，当两个以上的篇章单位形成修辞关系时，就构成了一种“树”结构——**修辞结构树**。篇章单位之间因存在某种修辞关系可以组合形成一个大的篇章单位，与相邻的篇章单位再构成更高层的修辞关系，继而得到所谓的**层次化篇章结构树**。每

个篇章的层次多少是不固定的,层次的多少由篇章中句子之间语义关系的复杂程度决定。通常情况下,语义关系越复杂,层次就越多。由于修辞关系含有特定语义,篇章结构关系也就表达了篇章内部的语义关系。

值得指出,相比 Hobbs 模型, RST 更关注篇章的整体性和连贯性,注重句子内部的篇章结构,篇章单位可以小到短语。修辞结构理论在篇章计算模型方面较受关注,它不仅作为篇章分析模型使用,也常常作为篇章生成模型使用^[19-20]。

1.2.1.4 宾州篇章树库体系

Marcu^[21-24]在修辞结构理论的基础上,对篇章结构关系问题进行了系统研究,为 Prasad 等^[14]构建宾州篇章树库(Penn Discourse Tree Bank, PDTB)提供了理论基础。相比修辞结构理论, PDTB 体系借鉴了谓词—论元表示形式,凸显了连接词的作用,其以连接词(谓词)为核心,标注与之相关的篇章单位(论元); PDTB 体系的篇章单位也与 RST 的略有不同,即其不再考虑短语级的语言单位作为篇章单位; PDTB 体系并不刻意构建篇章层次,但可以从相邻的篇章关系论元位置推导出部分篇章结构。PDTB 体系对篇章连接关系的两个论元,简单的记为 Arg1 和 Arg2。Arg2 指出现在从句中和连接词在句法上相邻的论元, Arg1 是另外一个论元。下面例子中, Arg1 用斜体表示, Arg2 用粗体表示,连接词用下划线标示。PDTB 体系中的篇章连接关系包括:

- (1) 显式关系(Explicit), 论元之间有明确的连接词, 如例 1.4:

例1.4 *The city's Campaign Finance Board has refused to pay Mr. Dinkins \$95,142 in matching funds* because **his campaign records are incomplete.**
(Contingency.Cause.Reason) (wsj_0041)

- (2) 隐式关系(Implicit), 在论元之间没有明确的连接词, 但篇章关系是可以推断出来的, 可以插入一个连接词, 句子仍然通顺。如例 1.5 中可以插入连接词“however”。

例1.5 *The city's Campaign Finance Board has refused to pay Mr. Dinkins \$95,142 in matching funds because his campaign records are incomplete.* Implicit = however
The campaign has blamed these reporting problems on computer errors.(Comparison.Contrast) (wsj_0041)

- (3) 替代关系(AltLex), 篇章关系是可以推断的, 但插入连接词会显得冗余。如例 1.6:

例1.6 *After trading at an average discount of more than 20% in late 1987 and part of last year, country funds currently trade at an average premium of 6%.* AltLex [The

reason:] **Share prices of many of these funds this year have climbed much more sharply than the foreign stocks they hold.** (wsj_0034)

(4) 实体关系 (EntRel), 论元之间没有篇章关系, Arg2 只是用于对 Arg1 中的某个实体提供进一步的描述, 如例 1.7:

例1.7 *Pierre Vinken will join the board as a nonexecutive director Nov. 29.* **EntRel**
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group. (wsj_0001)

(5) 无关系 (NoRel), Arg1 和 Arg2 之间没有篇章关系, 如例 1.8:

例1.8 *Mr. Rapanelli met in August with U.S. Assistant Treasury Secretary David Mulford.* **NoRel** **Argentine negotiator Carlos Carballo was in Washington and New York this week to meet with banks.** (wsj_0021)

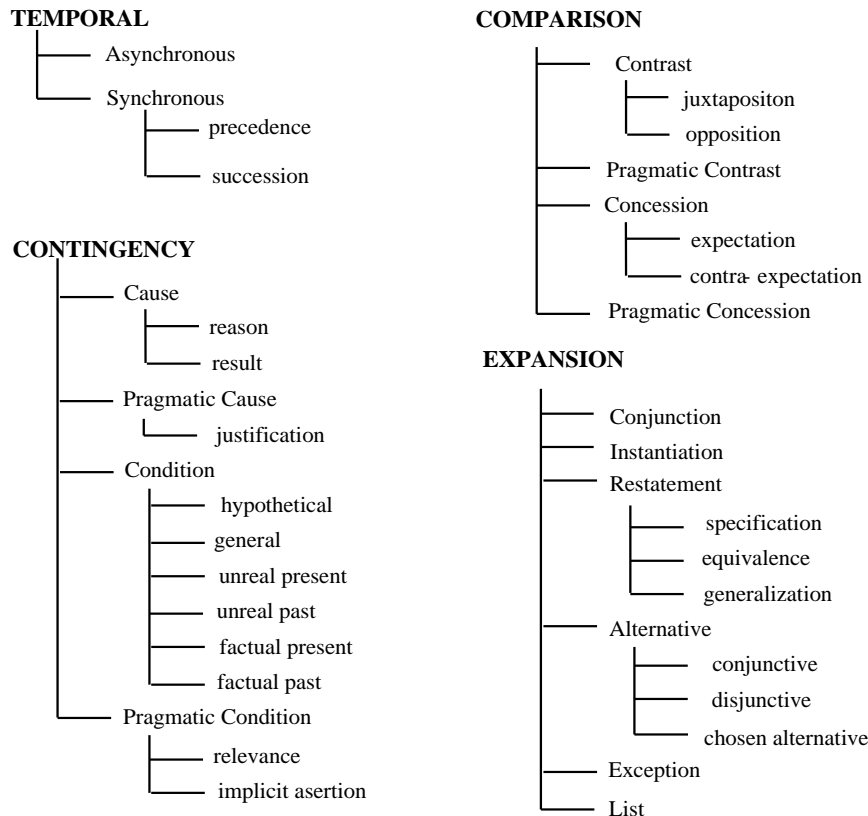


图1.5 PDTB 体系篇章语义关系层次图

PDTB 体系对每个显式和隐式关系都标注了篇章语义关系, 语义关系基本沿用修辞结构理论定义的 25 类篇章修辞关系, 所不同的是 PDTB 体系的语义信息采用层次化的方法来组织, 分别是: 类别、类型、子类。最顶层的类别代表主要的语义类别, 分为四类: TEMPORAL, CONTINGENCY, COMPARISON 和 EXPANSION。对每一个类别, 一系列的类型被定义出来细化该类别, 第二层 16 类。第三层的子类用来指定

论元的语义贡献，第三层 23 类（部分关系只到第二层），如例 1.4 为“Contingency.Cause.Reason”三层关系，例 1.5 为“Comparison.Contrast”二层关系。PDTB 体系语义关系层次如图 1.5 所示。

1.2.1.5 其它相关理论

除以上篇章结构理论外，还有其它英篇章章结构理论。Grosz 和 Sidner^[11]认为，篇章是具有意图的（因为人们写作本身就有某种意图）。因此，篇章结构理论不应只考虑篇章的内容，还应解释其中的意图。为此，她们提出了意图结构（intentional structure）作为篇章结构理论的基础。意图结构与修辞结构理论存在共同的基础。Moser 和 Moore^[25]认为，意图结构中的支配（dominance）和修辞结构理论中的核（nuclearity）相对应。除此之外，Asher 等^[26]和 Lascarides 等^[27]扩展了 Hobbs 模型中的篇章关系。Scha 和 Polanyi^[28]采取句法分析思想，为篇章结构提供了分析模型。Gardent^[29]提出一种基于特征的篇章树邻近文法（Feature-based Discourse Tree Adjoin Grammar）。基于此，Forbes^[30]又提出一种词汇化的篇章树邻近文法（Lexicalized Discourse Tree Adjoin Grammar）。Wolf 等^[31]研究表明基于树结构的篇章语料库在描述篇章结构时存在一些局限，不能有效表达子句间可能存在多种篇章关系情形。因此他们提出将篇章表示为一个链图，图中的弧可以有向的也可以是无向的，他们的语义关系大致遵循 Hobbs 的语义关系。

1.2.2 英语篇章结构分析的资源建设

现有资源主要包括修辞结构理论篇章树库（RSTDT）^[32-33]、宾州篇章树库（PDTB）^[15]等。下面对这两种资源标注情况进行详细阐述。

1.2.2.1 修辞结构理论篇章树库

英语 RSTDT^[32]以 RST 为支撑，构建篇章标注语料库（RST Discourse Treebank, RSTDT）。所标注的 385 篇华尔街日报文章皆取自宾州树库，篇幅长度不等，从 31 个词到 2124 个词，总词数达到 176000，平均每篇文章 458 个词。文章的内容涉及到各种话题，如财政报道、商业新闻、文化点评、编者按、读者来信等。

篇章结构分析的首要任务是确定基本篇章单位（Elementary Discourse Unit, EDU）。在具体标注时，研究者对基本篇章单位的规定如下：充当主语或宾语的从句不属于基本篇章单位；充当主要动词补语的从句不属于基本篇章单位；所有词汇或句法标记的

起状语作用的从句都属于基本篇章单位,包括起状语作用的非谓语动词词组;定语从句、后置的名词修饰短语或将其它基本篇章单位割裂开的从句或非谓语动词短语为基本篇章单位;有明显篇章标记的短语作为基本篇章单位,如因为(because),尽管(in spite of),根据(according to)等引导的短语是基本篇章单位。

基本篇章单位确定下来后,剩下的工作就是根据 RST 确定基本篇章单位之间的关系,进而生成有层次的篇章结构树。篇章关系可能是单核关系或多核关系。文献[8]提出 RST 时只给出 20 多种修辞关系,但他们明确指出这是一个开放关系集,这意味着读者可以在给定话语的内部定义出其它的关系类型。RST 篇章树库中标注了 53 种单核心关系和 25 种多核心关系,78 种关系又分成 16 个组别,每组都具有相同的修辞功能,如:原因、比较、对比、条件等。RST 中的关系分类如表 1.1 所示。

表1.1 RSTDT 中部分关系及中心情况表

Type	Mononuclear(satellite)	Mononuclear(nucleus)	Multinuclear
Background	Background, circumstance		
Cause	Consequence	Cause, Consequence	cause-result, Consequence
Comparison	Comparison, preference, analogy		Comparison, analogy, proportion
Condition	Condition, hypothetical, contingency, otherwise		otherwise
Explanation	Evidence, explanation-argumentative, reason		reason
Joint			List, disjunction
Summary	Summary, restatement	summary	
Temporal		temporal-before, temporal-after, temporal-same-time	Sequence, inverted sequence

为了制定高质量的前后一致的标注标准和方法,研究者采用人工标注的方法。他们所选用的标注者都是有过标注经历的、从事篇章分析和新闻报道的专业人员。在正式标注前,标注者接受了篇章结构标注培训。在整个语料库的建设过程中,研究者一直设法保证标注者之间内部的一致性。例 1.9 的 RST 标注保存形式如图 1.6 所示,例 1.9 中有 3 个基本篇章单位(例中分别用数字标示),其中 2 和 3 是“elaboration-additional”,2 是核心。

例1.9 1 Spencer J. Volk, president and chief operating officer of this consumer and industrial products company, was elected a director. 2 Mr. Volk, 55 years old, succeeds Duncan Dwight, 3 who retired in September. (wsj_0600)

```
( Root (span 1 3) (prom 1)
  ( Nucleus (leaf 1) (rel2par span) (prom 1) (text Spencer J. Volk, president and chief operating
officer of this consumer and industrial products company, was elected a director.) )
  ( Satellite (span 2 3) (rel2par elaboration-additional) (prom 2)
    ( Nucleus (leaf 2) (rel2par span) (prom 2) (text Mr. Volk, 55 years old, succeeds Duncan
Dwight,) )
    ( Satellite (leaf 3) (rel2par elaboration-additional-e) (prom 3) (text who retired in
September.) )
  )
)
```

图1.6 例 1.9 在 RSTDT 中的保存结果

1.2.2.2 宾州篇章树库

PDTB 是一个在宾夕法尼亚大学研究的美国自然基金资助项目。该项目的目标是开发一个标注有篇章结构信息的大规模语料库，主要标注与篇章连接词（discourse connectives）相关的连贯性关系（coherence relation）。标注信息主要包括连接词的论元结构、语义区分信息，以及连接词和论元的属性等相关特征。该目标注了 100 万字的 Treebank-2 语料库(LDC95T7)中所提供的篇章关系，目前的版本是 PDTB2.0^[15]。PDTB2.0 标注了以下几种论元结构关系：显式关系（Explicit）、隐式关系（Implicit）、替代关系（AltLex）、实体关系（EntRel）、无关系（NoRel）。除了论元结构关系，PDTB2.0 中显式关系、隐式关系、替代关系都标有语义类别信息和属性信息。虽然篇章可以从多个方面进行描述，PDTB 专注于描述篇章关系。PDTB 的标注遵从词汇化方法，采用与理论无关的做法，目的是为了语料在不同的理论框架内使用。

PDTB 基于华尔街日报，共标注了 2304 个文档，约 100 万词，其中以新闻为主。PDTB2.0 在 2008 年 2 月发布使用。PDTB2.0 总共有 25 部分，建议使用 2-21 部分做训练集，22 部分作为开发集，23 部分作为测试集。0,1,24 如果需要的话可以作为附加的开发集。PDTB2.0 中共标注了 40600 个关系，其中 18439 个显式篇章关系，16224 个隐式篇章关系，624 个由非连接词表示的篇章关系，5210 个通过实体重复或共指表示的关系，还有 254 个相邻句子不存在所定义的关系。表 1.2 给出了每种关系的具体分布情况。

表1.2 PDTB2.0 关系分布情况

SECTION	Explicit	Implicit	AltLex	EntRel	NoRel	TOTAL
Sec.00	712	592	19	218	30	1571
Sec.01	750	591	20	271	3	1635
Sec.02	713	708	28	200	8	1657
Sec.03	529	446	13	166	2	1156
Sec.04	822	747	27	238	5	1839
Sec.05	816	780	29	148	15	1788
Sec.06	653	571	14	143	13	1394
Sec.07	804	676	24	302	5	1811
Sec.08	176	161	2	56	8	403
Sec.09	786	720	13	193	31	1743
Sec.10	720	613	5	185	15	1538
Sec.11	780	839	35	208	16	1878
Sec.12	800	726	35	180	16	1757
Sec.13	941	863	32	250	8	2094
Sec.14	734	731	31	244	13	1753
Sec.15	868	703	40	201	5	1817
Sec.16	1092	993	61	243	5	2394
Sec.17	614	487	22	201	6	1330
Sec.18	898	722	32	265	7	1924
Sec.19	647	519	34	250	7	1457
Sec.20	724	627	20	257	14	1642
Sec.21	605	524	27	203	5	1364
Sec.22	680	522	19	215	8	1444
Sec.23	923	769	30	217	4	1943
Sec.24	672	423	12	156	5	1268
ALL	18459	16053	624	5210	254	40600

PDTB 对显式连接关系、隐式连接关系和替代关系都标注了篇章语义关系类别信息，具体篇章语义关系类别见图 1.5。与动词类似，篇章连接词同样具有不止一个语义，具体语义依赖于篇章上下文和论元的内容。例如：since 有三种语义，可以是纯粹的时间，可以是纯粹的原因，还可以是同时表示原因和时间，since 的三种语义分别如例 1.10 中 A、B 和 C 所示。

例1.10 A) The Mountain View, Calif., company has been receiving it was demon-1,000 calls aday about the product since strated at a computer publishing conference several weeks ago.

B) It was a far safer deal for lenders since NWA had a healthier cash flow and more collateral on hand.

C) Domestic car sales have plunged 19% since the Big Three ended many of their programs Sept. 30.

语义标注的目标主要是区分连接词到底是哪种语义，根据语义可知连接词所连接

的论元之间的语义关系，当论元之间有两个以上语义关系时，多种语义关系均被标注。

PDTB 对每种关系都存储了关系类别、关系的位置、连接词、语义类别关系、论元内容及位置、文本内容等信息。例 1.4 在语料中的保存结果如图 1.7 所示。保存信息首先以“___”标示，如“___Explicit___”、“___Arg1___”，然后给出相应信息的位置信息，用位置和 Gorn addresses 两种形式表示，“#### Text ####”后标示文本内容，“#### Features ####”后给出相应的特征标注，显式、隐式和替代关系根据图 1.5 给出的 PDTB 篇章语义类别具体标注。

___Explicit___
3317..3324
29,1,1,2,0
Text
because
#####
Features
Wr, Comm, Null, Null
because, Contingency.Cause.Reason
___Arg1___
3226..3316
29,0;29,1,0;29,1,1,0;29,1,1,1;29,2
Text
The city's Campaign Finance Board has refused to pay Mr. Dinkins \$95,142 in matching funds
#####
Features
Inh, Null, Null, Null
___Arg2___
3325..3360
29,1,1,2,1
Text
his campaign records are incomplete
#####
Features
Inh, Null, Null, Null

图1.7 例 1.4 Explicit 篇章关系保存实例

1.2.3 英语篇章结构分析的计算模型

相关计算模型通常基于特定的资源开展，下面首先介绍常用的评测方法，然后介绍基于 RSTDT 和 PDTB 的研究。

篇章结构分析评测主要考虑算法的**正确率**和**F1 值**两个性能指标。其中，正确率（Accuracy）采用公式 1.1 进行度量。公式 1.1 中 TruePositive 代表本来是正样例，同时被分类成正样例的个数；TrueNegative 代表本来是负样例，同时被分类成负样例的个数；All 代表需要分类的样例总个数。

$$\text{Accuracy} = \frac{\text{TruePositive} + \text{TrueNegative}}{\text{ALL}} \quad \text{公式 1.1}$$

F 值由准确率（Precision）和召回率（Recall）共同体现。召回率（Recall）是结果中正确的对象数目占总对象数的百分比，它反映的是系统的完备性。准确率（Precision）是结果中正确的对象数目占实际对象数目的百分比，它反映了系统的准确程度。通常采用这两个指标的综合值—F 值，准确率（Precision），召回率（Recall）和 F 值公式如下：

$$\text{Precision} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad \text{公式 1.2}$$

$$\text{Recall} = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad \text{公式 1.3}$$

$$F = \frac{(\beta + 1)P \times R}{P + R} \quad \text{公式 1.4}$$

公式 1.2 和 1.3 中，FalsePositive 代表本来是负样例，但被分类成正样例的个数（通常叫**误报**）；FalseNegative 代表本来是正样例，但被分类成负样例的个数（通常叫**漏报**）。公式 1.4 中， β 为召回率和准确率的相对权重，一般取 1，此时准确率（Precision）和召回率（Recall）的综合值为 F1。

1.2.3.1 基于 RSTDT 的研究

在修辞结构理论篇章树库（RSTDT）上进行篇章结构分析研究一般分为基本篇章单位（Elementary Discourse Units, EDU）识别和篇章结构生成两步。

关于 EDU 的自动识别研究较多，结果也比较理想。其中比较有代表性的研究包括：Soricut 等^[19]采用基于统计的方法进行识别，EDU 识别在自动句法树上获得 F1 值 **83.1%**，在标准句法树上 F1 值为 84.7%。Hernault 等^[34]给出了一个基于序列数据标注的篇章分割模型，使用词汇和句法特征，采用 **CRF**，实验表明作者的序列篇章分割模型 F1 值达到 **94%**，接近于人工篇章分割的 F1 值 98%。由上可知，目前 RSTDT 上 EDU 识别准确率较高，进一步提升的空间不大。

篇章结构生成方面，结果则不理想。Soricut 等^[19]利用语法和词法信息进行句子

级的篇章结构分析, 他们的算法称为 **SPADE**, 在篇章关系识别时采用概率模型计算各种篇章关系的概率。篇章结构分析模型采用全自动的方法, 识别无标注篇章关系的 F1 值为 **70.5%**, 采用标准的基本篇章单位和标准句法树的结果是 **96.2%**。但 **SPADE** 并不对整篇文本进行篇章关系识别, 作者通过实验证明句法和篇章信息的关系, 在识别时作者没有用到线索词。**Huong** 等^[35]研究书面文本篇章结构生成, 给出了一个书面文本自动篇章结构生成系统, 系统分为两个层次: 句子级的篇章结构分析和文本级的篇章结构分析。句子级的篇章结构分析使用句法和线索词来进行基本篇章单位识别和篇章结构生成。对于篇章级别, 为降低篇章结构分析的搜索空间, 加入了文本相邻和文本组织限制。实验采用 **RSTDT**, 结果表明降低搜索空间后 F1 值为 70.1%, 缺点是计算量较大。**David** 等^[36]提出了一个基于支持向量机的篇章结构分析器, 作者给出了一种新的方法进行篇章结构分析, 采用统计机器学习的方法, 使用了丰富的特征空间, 给出的是整篇文档的篇章结构树, 不仅仅是句内结构树。作者并没有进行 **EDU** 识别, 而是采用现成的工具进行 **EDU** 识别。**Hernault** 等^[37]在 **RST** 上实现了基于 **SVM** 的篇章结构分析器 **HILDA**。对篇章切分和关系识别使用 **SVM** 训练了分类器, 采用贪婪的自底向上的方法构建篇章结构树, 篇章结构树构建的时间复杂度取决于输入文本的长度。**HILDA** 可以对所有文本构建整棵篇章结构树, 然而 **SPADE** 只能进行句内分析。**HILDA** 在树构建和篇章关系分析上的效果较好, 结构识别 F1 值为 72.3%, 完整句法树识别 F1 值为 44.3%。**Feng** ^[38]在 **HILDA** 的基础上进行篇章结构树构建和关系识别, 作者抽取了更丰富的特征, 性能比 **HILDA** 有所提升, 另外, 作者分析了句内、句间和综合情况下的篇章结构分析性能。**Joty** 等^[39]给出一种使用动态条件随机场进行句子级的篇章分析方法, 使用人工 **EDU** 切分结果识别 18 类关系 F1 值为 77.1%。**Joty** 等^[40]将篇章结构分析分为句子级和篇章级采用 **CRF** 模型分别进行, 然后采用两种方法将篇章级和句子级的模型融合进行篇章结构分析, 得到的结果较 **HILDA** 有较大提高, 结构识别 F1 值为 **82.5%**, 关系识别 F1 值为 **55.7%**。由于 **Joty** 等^[40]的时间复杂度较高, **Feng** 等^[41]给出一种线性时间的篇章结构分析方法, 利用两个 **CRF** 模型, 采用加入约束和后编辑的自底向上方法构建篇章结构树, 篇章分析速度比 **Joty** 等^[40]快, 结构识别 F1 为 **85.7%**, 关系识别 F1 值为 **58.2%**。此外, **Ji** 和 **Eisenstein**^[42]参考 **DeepLearning** 做法, 采用线性变换将表面特征转化成隐式空间进行基于移进规约的篇章级结构分析方法。可见, 篇章结构分析中的整体结构树构建是目前 **RSTDT** 上的研究热点和难点。

1.2.3.2 基于 PDTB 的研究

宾州篇章语料库 (PDTB) 的构建显著推动了篇章结构分析的研究, 在篇章计算方面受到了极大的关注。目前的研究主要集中在论元识别和篇章关系识别。

论元识别方面, Wellner 等^[43]自动识别篇章中的论元, 在标准句法树上获得 74.2% 的正确率, 在自动句法树上获得 64.6% 的正确率。Elwell 和 Baldridge^[44]自动进行篇章连接词及论元识别, 在 Wellner 等^[43]研究基础上, 添加了形态学、句法、词汇、篇章模式等特征, 在连接词识别、论元 Arg1 和论元 Arg2 识别方面都获得了较好的结果, 在标准句法树上正确率分别达到 77.8%、82% 和 93.7%, 在自动句法树上正确率分别是 73.6%、80% 和 90.2%。Wellner^[45]用序列模型和排序方法进行篇章分析, 在标准句法树上 Arg1 识别正确率是 80%, Arg2 识别正确率是 91%, 在自动句法树上 Arg1 识别正确率是 62%, Arg2 识别正确率是 86%。上述作者在识别论元时均采用论元中心代替整个论元, 使得识别相对简单。Prasad 等^[46]利用范围学习进行浅层篇章结构分析, 重点识别 Arg1, 主要处理 Arg1 和 Arg2 在不同句子中的情况, 识别的结果是包含 Arg1 的句子。实验表明利用范围学习的 Arg1 识别效果有较大的提升, 正确率达到 86.3%。不过需要指出的是, 作者识别的是包含论元的句子, 并不是真正的论元。徐凡^[47]针对 PDTB 的篇章论元识别问题, 分别从句内 (篇章连接词与论元处于同一句) 和句外 (篇章连接词与论元不处于同一句) 两种情况加以处理。对于句内情况, 分别提出了基于浅层语义分析和基于句法树裁减的两种方法, 不同于现有的基于组块的方法, 他提出的方法在识别篇章单元层面从单词过渡到组块级别, 同时本文采用了句法树中更多的结构化信息。对于句外情况, 提出了一种轻量级基于规则的解决方案。实验结果表明, 句内 Arg1 精确匹配 F1 值为 79.38%, Arg2 精确匹配 F1 值为 91.38%, Both 精确匹配性能为 74.87%; 句外 Arg1 精确匹配性能为 39.19% 的 F1 值, Arg2 精确匹配性能为 82.74%, Both 精确匹配性能为 34.19%; 对于整体情况 (句内和句间), Arg1 精确匹配性能为 63.76% 的 F1 值, Arg2 精确匹配性能为 88.02%, Both 精确匹配性能为 59.06%。

篇章关系识别方面, Dinesh 等^[48]针对 PDTB 篇章语料库中表示连接方式的某些词作了初步消歧研究。Pitler 等^[49]指出, 在 PDTB 篇章语料库中隐式篇章关系与显式篇章关系大约各占一半。由于显式篇章关系中连接词 (connective) 的存在且歧义较少 (大约只有 2%), 因此比较容易识别。这使得隐式篇章关系研究成为篇章结构关系分析成败的关键。基于此, Pitler 等^[50]深入研究了不同类型语言特征对隐式篇章关系识别的贡献, 实验发现情感倾向标志、动词类别、动词短语长度、情态动词、上下文

环境和词法等特征对篇章关系识别具有一定作用，在 PDTB 上正确率为 44.58%（6 大类，4 大类加实体关系和无关系）。Lin 等^[51]针对 PDTB 篇章语料库的特点，把篇章关系连接语作为谓词，把谓词管辖的文本单元（如句子、子句、从句等）作为论元，探索了各种上下文特征、词对特征、成分句法信息和依存句法信息，对隐式篇章关系进行了研究，在 PDTB 上正确率为 40.2%（二层出现次数较多的 11 类关系）。相对于 Pitler 等^[50]，Lin 等^[51]不再将词分类，而是完全依据统计信息捕捉词对特征，降低了数据处理复杂性，另外引入了上下文信息、成分句法信息和依存句法信息，开拓了隐式篇章关系研究的新路径。Wang 等^[52]试图对 Pitler^[50]等和 Lin 等的研究思路进一步拓展，针对 Lin 等只利用二层句法结构信息作为特征因而获取结构化信息量较少的缺点，采用卷积树核函数将句法结构信息获取能力扩大到整个句法树的多层结构，并汲取了 Pitler 等^[50]利用词类信息的思路，捕捉文本中对识别篇章关系贡献较大的时序信息，摒弃了对识别篇章关系贡献较小的多数词类信息。不过实验结果并不理想，在宾州篇章树库上的隐式篇章关系识别正确率只有 38.4%（6 大类）。徐凡等^[53]系统地探索了篇章中的浅层语义信息和以态度韵为导向的句子级情感等平面特征对隐式篇章关系识别的作用，同时提出了一种简单而有效的树核方法，最后采用复合核方法加以集成。在 PDTB2.0 语料库上的实验结果表明，引入浅层语义和情感等信息后，正确率得到显著提升，使用平面特征正确率为 52.0%，使用复合核函数的方法可以达到 53.1%。Fisher 和 Simmons^[54]采用半监督的方法进行 11 类隐式关系类别，使用有标签的数据 F1 值达 41.1%。

在端到端的篇章结构分析方面，Lin^[55]研究如何在 PDTB 上进行篇章结构分析，对于难度较大的隐式篇章关系识别，采用上下文、词对、句法特征、依存树特征进行识别。整个系统包括连接词识别、论元识别、显式关系分类、隐式关系分类、属性标注，这是第一个端到端的 PDTB 分析工作。徐凡^[47]在英语 PDTB 语料上进行了篇章结构分析关键问题研究，实现了一个全新的基于树核的英文篇章结构分析平台，不同于已有的基于特征向量的篇章结构分析平台，他采用了树核方法，能够有效结合文本中的平面信息和结构化信息，从而性能上要优于已有的方法。

需要指出的是，由于在语料库使用上存在一些差别，上述方法的优劣并不能完全依据实验提供的正确率进行判断。不过，我们依然可以看出，在二层主要关系类别上，相对显式篇章关系 90% 以上的识别正确率^[49]，隐式篇章关系的识别正确率徘徊在 50% 左右，隐式篇章关系识别成为篇章结构分析成败的关键。

1.2.3.3 结合 RSTDT 和 PDTB 的研究

Hernault 等^[56]提出利用特征向量扩充提高少量篇章关系分类的一种半监督方法。许多篇章结构分析采用全监督的机器学习方法，此类方法要求标注训练语料，费时且代价高，而没有标注信息的数据则较为容易大量获得。作者提出了一种半监督方法，采用以前使用过且报告有用的特征进行篇章关系学习，探索在无标注的文档上进行篇章关系识别的可能性，特别是改进类别例子较少的篇章关系性能。作者采用同现矩阵进行特征扩充，在 RSTDT 和 PDTB 上都进行了实验，结果表明 Accuracy 和 F1 值都有较大的提高。文章首次提出关于出现频率较少的篇章关系的处理方法，这种方法对语料较少的领域也比较有用。

上述研究主要是采用半监督的方法进行篇章关系的识别，解决了某些领域篇章标注较少的情况，但方法的性能有待进一步的提升。

1.2.4 汉语篇章结构分析研究现状及存在问题

1.2.3 节主要介绍了英语篇章结构分析的研究现状，汉语篇章结构分析起步较晚，目前的研究成果不多，下面介绍汉语篇章结构分析相关研究情况。

1.2.4.1 汉语篇章结构分析研究现状

汉语篇章分析的兴起和发展大致经历了四个阶段（郑贵友^[57]）：第一个阶段纯粹以文章写作为主要目的，对篇章构成加以观察；第二个阶段以文章学分析为主，同时从语言学的角度对篇章构成加以观察；第三个阶段从语言学的角度，观察汉语篇章结构规律，具有“本土特征”，此时语言学家们更多地关注了汉语篇章结构的“本土特征”一句群，确立了句群作为汉语篇章观察研究“标本”的地位，显著加强了汉语篇章内部微观语义结构、篇章内部衔接手段的研究（吕叔湘^[58]；曹政^[59]；吴为章等^[60]），同时汉语复句研究也得到了较大发展；第四个阶段引进西方现代篇章语言学理论，研究汉语篇章问题，比较有影响的是采用 RST 和 PDTB 体系。

相比英语，汉语篇章结构研究刚刚起步。资源构建主要采用三种方法：

1) 基于 RST 的标注。乐明^[61-62]和陈莉萍^[63]均进行了基于修辞结构理论的汉语篇章结构语料库研究，但他们的具体工作实现有较大差异。乐明^[61-62]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。他定义了 12 类 47 种汉语修辞关系，以句号、问号、叹号、分号、冒号、破折号、省略号及段落结束符等为标

记定义汉语基本篇章单位，完成 97 篇财经评论文章（来自中国大陆主要媒体）的修辞结构标注。陈莉萍^[63]试图采用 RST 标注汉语篇章，其基本篇章单位以标点分割，如“目前，……”中的“目前”也会作为基本篇章单位。他们的研究都表明 RST 的很多篇章关系无法在汉语中找到与之对应的关系。综上，汉语基本篇章单位和英语基本篇章单位有差别，乐明对基本篇章单位的定义显然太粗，陈莉萍的定义显然太细。

2) 基于 PDTB 体系的标注。Xue^[64]和 Zhou 等^[65]尝试使用 PDTB 体系标注汉语，PDTB 体系以连接词为谓词标注其论元结构。但汉语中连接词大量缺省，PDTB 体系表现出很大不适应；又由于连接词并不能覆盖每一个篇章单位，PDTB 体系通常不能构建一个完整的篇章结构，这对于篇章结构分析而言显然缺少了很重要的内容。Huang 和 Chen^[66-67]对汉语篇章结构进行了系列研究，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，构建了一个篇章语料库，完成了 3081 个句对的小规模的中文篇章树库，在标注过程中简单的把句子作为一个论元，标注了四种关系（Temporal, Contingency, Comparison, Expansion）。Huang 和 Chen^[68]提出一种汉语篇章结构标注框架，并开发了一套网上标注系统，在此标注框架下，标注者选择出显式或隐式、句内或跨句等篇章关系，然后按照 PDTB 的关系分类体系进行标注，另外，作者还标注了每个实例的情感信息备用。Zhou 等^[69]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^[70]在英文篇章关系研究的基础上分析了中英文的差异，总结了汉语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，整体上还是参照英语的 PDTB 体系进行标注，目前发布了哈工大中文篇章关系语料（HIT-CDTB）²。HIT-CDTB 包括 525 篇标注文本，语料生文本来源于 OntoNotes 4.0，覆盖了句群关系、复句关系、分句关系等多级信息。标注采用宾州篇章树库的模式。标注的关联词分为显式关联词和隐式关联词两种。显式关联词包括普通关联词（但是，由于等）、带修饰关联词（部分原因，尤其是等）和平行关联词（一方面…另一方面…，一边…一边…等）。标注的篇章关系共 6 个大类：时序关系、因果关系、条件关系、比较关系、扩展关系和并列关系。由于很多时间词（如九八年）被标注为连接词，所以语料中标注出来的显式关联词共有 1472 种，它们总共出现的次数是 11519 次。

3) 采用汉语本土理论标注。参考邢福义^[71]的汉语复句研究成果，华中师范大学

² <http://ir.hit.edu.cn/hit-cdtb/>

标注了汉语复句语料库 (the Corpus of Chinese Compound Sentences)³, 目前已收有标复句 658447 句, 约 44395000 字, 语料来源以《人民日报》和《长江日报》为主。但汉语有标复句只占汉语复句的 30% 左右^[72], 这就使得该语料库的应用受到很大限制。而且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, 这显然不能满足篇章结构分析的需求。清华汉语树库 (Tsinghua Chinese Treebank, TCT)^[73]是从大规模的经过基本信息标注 (分词和词性标注) 的汉语平衡语料库中, 提取出 100 万汉字规模的语料文本, 经过自动断句、自动句法分析和人工校对, 形成的高质量汉语句法树库语料。TCT 中标出了复句内各分句之间的关系信息, 复句分类采用比较常用的并列关系、连贯关系、递进关系、选择关系、因果关系、目的关系、假设关系、条件关系、转折关系分类方法。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。

由于汉语篇章结构语料缺乏, 汉语篇章结构分析研究受到了制约。Xue 等^[74]和 Yang 等^[75]采用基于逗号的篇章分析方法切分汉语句子, 值得指出, 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, 无法准确反映实际情况。张牧宇等^[76]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, 并给出初步的实验结果, 但其所标语料参考英语 PDTB 体系, 不能进行完全的篇章结构分析, 只能进行部分篇章分析。涂眉^[77]等在 TCT 上进行了基于最大熵的汉语篇章结构自动分析方法, 实验结果表明, 篇章语义单元自动切分的 F1 值能达到 89.1%, 当篇章语义结构树高度不超过 6 层时, 篇章语义关系标注的 F1 值为 63%。

汉语篇章分析有多个方面, 本文主要讲述篇章级结构分析, 目前汉语话题之间也有相关的研究, 篇章结构和话题之间存在联系: 一方面, 一个话题的若干个单元之间应该可以组成一个完整的篇章结构树; 另一方面, 从篇章结构树上可以体现话题。目前汉语话题理论方面比较有代表性的是宋柔等^[78]提出的广义话题理论, 根据汉语篇章的特点, 以标点句为基础, 提出广义话题和话题句的概念, 描述了汉语的话题结构和话题句特征, 给出了话题句动态堆栈模型。目前已发布《鹿鼎记》第一回、《围城》全书和其它语料共约 40 万字。

³ <http://ling.ccnu.edu.cn:8089/jiansuo/TestFuju.jsp>

1.2.4.2 存在问题及研究趋势

适用于汉语篇章结构分析的理论还不完善。适用于英语篇章结构分析的修辞结构理论和宾州篇章树库体系都不太适合汉语本身的特点。乐明^[61]采用修辞结构理论标注汉语篇章,发现修辞结构理论的很多篇章关系在汉语中无法找到对应关系,而且汉语的基本篇章单位和英语的基本篇章单位存在较大差别。Xue^[65]尝试使用宾州篇章树库体系标注汉语,但宾州篇章树库的关系分类体系也不太适合汉语,而且关系标注中没有区分论元的主次,这部分信息对理解篇章及相关自然语言处理应用都非常重要。目前,有关汉语篇章的相关理论研究主要在复句和句群上,虽然复句和句群是篇章的重要部分,但复句和句群理论研究并不能代替篇章结构理论研究。

符合汉语特点的大规模汉语篇章结构资源匮乏。乐明^[61]采用了修辞结构理论标注汉语财经篇章,但其基本篇章单位的分割不适合汉语,对复句内部的关系也没有标注。哈工大中文篇章关系语料(HIT-CDTB)采用PDTB体系,其所标关系也不完全适用于汉语,标注只考虑篇章连接词及其论元,没有考虑篇章层次结构及中心信息。汉语复句语料库只考虑有标复句,并且只考虑复句内部关系,没有考虑句子和句子之间的关系。据姚双云^[72]统计,《人民日报》语料样本583,181个复句中,有标复句有165,096句,占复句总数的28.3%,无标复句有418,085句,占复句的71.7%。可见无标复句所占的比例较大,篇章结构分析的难点是处理无标篇章结构关系。

汉语篇章结构分析的计算模型研究缺乏。由于适用于汉语篇章结构分析的理论还不完善,并且缺乏大规模汉语篇章结构资源的支持,目前相关有针对性的计算模型研究较少。基于财经篇章语料库的研究^[61]和基于华中师大复句语料库的研究^[72]主要还是根据标注语料进行语言学统计分析,并没有进行有效的篇章语义计算。张牧宇等^[70]HIT-CDTB研究还很初步,需要更深入的计算分析研究。

汉语篇章结构分析起步较晚,目前的研究较少,针对汉语的篇章结构语料匮乏,需要首先标注汉语篇章结构语料库,随着语料库的逐步完善,汉语篇章结构分析的研究将迎来一个新的阶段。在汉语篇章结构分析中,可以参考英语篇章结构分析的结果,采用类似的方法进行,但由于汉英不同,需要探索针对汉语的篇章结构分析方法。

1.3 本文的研究内容

针对汉语篇章结构分析研究存在的问题,本文研究内容主要包含以下三个方面:
一是参考RST和PDTB体系。结合汉语特点,提出**基于连接依存树的汉语篇章**

结构表示体系，并给出连接词、子句、篇章结构关系及篇章单位主次的定义及判断方法，其中连接依存树的内部节点为连接词，连接词既可以表示篇章结构的层次，也可以表示篇章逻辑语义关系，连接词所连接的篇章单位之间有主次之分。

二是研究汉语篇章结构语料库的构建。不同于英语的方法，在连接依存树表示体系的指导下，采用自顶向下的标注策略和人机结合的语料库标注方法构建汉语篇章结构语料库，并对标注语料库进行一致性测试和统计分析。

三是根据标注汉语篇章结构语料，设计并实现汉语篇章结构分析平台。汉语篇章结构分析包括**子句识别、关系识别、篇章结构树构建等基本任务**，本文首先分别研究单个任务，进而综合各个任务给出完整的汉语篇章结构分析平台。

1.4 本文的组织结构

本文主要进行汉语篇章结构表示体系及资源构建研究，组织结构如下：

第1章，绪论。首先介绍了本文的研究背景和研究意义，然后对中英文篇章结构分析的理论研究、资源建设和计算模型的国内外研究现状进行介绍，指出了目前存在的问题和不足，以及研究热点和趋势，最后给出了本文的研究内容。

第2章，研究汉语篇章结构的理论表示。参考 RST 和 PDTB 体系，本文提出了基于连接依存树的汉语篇章结构表示体系，并详细的介绍了连接依存树表示体系的几个关键元素子句、连接词、篇章结构关系、篇章单位主次的定义及判定方法。最后，将本文所提出的基于连接依存树的篇章结构表示体系和相关理论进行对比。

第3章，构建基于连接依存树表示体系的汉语篇章结构语料库(Chinese Discourse Tree Bank, CDTB)。首先介绍了语料库建设的一般情况，**提出 CDTB 标注采用自顶向下的语料库标注策略和人机结合的语料库标注方法**，给出了具体的标注流程和语料格式，然后进行了语料标注一致性实验，最后对 CDTB 语料标注情况进行了详细的统计分析。

第4章，介绍基于 CDTB 的汉语篇章结构分析。首先根据语料标注结果和相关研究成果给出汉语篇章结构分析平台基本框架，然后介绍了实验所用的特征和具体的实验设置，最后分别对**基于标点的子句识别、连接词识别与分类、隐式篇章关系识别、篇章单位主次识别和基于连接依存树的汉语篇章结构平台性能进行实验和详细分析**。

第5章，总结与展望。总结了本文取得的研究成果，同时也指出了存在的不足，并对未来的工作进行了展望。

第2章 基于连接依存树的汉语篇章结构表示体系

由第1章可知,目前汉语篇章结构分析的理论研究较少,因此需要建立适合汉语特点的篇章结构表示体系。参考已有研究,结合汉语特点,本章提出一种基于连接依存树的汉语篇章结构表示体系。首先介绍本文提出的连接依存树,然后重点介绍连接依存树的子句、连接词、篇章结构关系和篇章单位主次的特点及判定方法,最后将基于连接依存树的汉语篇章结构表示体系和相关篇章结构分析理论进行对比分析。

2.1 引言

2.1.1 已有篇章结构理论体系分析

一般认为,完整的篇章结构包括篇章单位、连接词、篇章结构、篇章关系、篇章主次。然而,由1.2节可知,已有的篇章结构理论并没有完整的展现这些信息。例如,修辞结构理论(RST)将篇章结构表示成由基本篇章单位(EDUs)层次组合而成的树状形式,然而,RST忽视了连接词对篇章结构的重要性。如例2.1.1用RST形式表示如图2.1所示,例2.1.1共有4个基本篇章单位,其中e1和e2是“attribution”关系,e1是核心单位,e2是卫星成分。e1和e2的组合和e3是“same-unit”关系,都是核心单位,在图2.1中用箭头指向的内容为中心。RST体现了e1-e3和e4是条件关系,但没有体现连接词“if”的信息。宾州篇章树库体系(PDTB)采用谓词论元的篇章结构表示形式,将连接词作为谓词,其所管辖的两个篇章单位称为论元。如例2.1.2中连接词是“particular if”,连接词所管辖的Arg1为“Catching up with commercial competitors in retail banking and financial services”,Arg2为“market conditions turn sour”,连接词的语义关系为“Contingency.Condition.Hypothetica”。和例2.1.1同样的内容,PDTB体系体现了连接词、论元(篇章单位)、篇章关系信息,但没有体现论元Arg1和Arg2的主次地位。

例2.1.1 [Catching up with commercial competitors in retail banking and financial services,] e1 [they argue,] e2 [will be difficult,] e3 [particularly if market conditions turn sour.] e4 (RST理论表示) (wsj_0616)

例2.1.2 [Catching up with commercial competitors in retail banking and financial services] Arg1, they argue, will be difficult, particularly if [market conditions turn sour] Arg2.

(Contingency.Condition.Hypothetical) (PDTB 体系表示)

(wsj_0616)

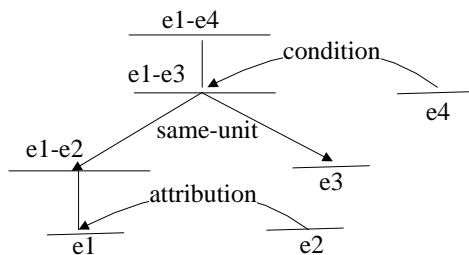


图2.1 例 2.1.1 RST 结构实例

显然，RST 和 PDTB 在表示篇章结构时有它们各自的优点，同时也存在一些问题：1) RST 注重层次结构，忽略了连接词的结构表达作用，但连接词对于篇章结构和篇章关系的作用都至关重要；2) PDTB 重视连接词的篇章结构作用，以其为基础构建篇章关系及论元，但由于连接词难以贯穿整个篇章，所以往往难以构建出整个篇章的结构。另外，以往这些理论主要在英语上实践，一些已有的研究表明，这些理论实践于汉语时在基本篇章单位定义、关系类别体系等方面都有一些不适应^[61, 63, 65]。

2.1.2 汉语篇章结构的特点

汉语在篇章结构上和英语存在差异，主要表现在以下几个方面：1) 汉语省略较多。汉语重“意合”，只要表达意义清楚，句子成分能省则省。而英语重“形合”，它强调句式结构完整，除省略句外，每个句子必须有主语^[79]。2) 汉英连接方式差异。汉语较少借用形式连接手段，篇章主要是通过逻辑联系或语序、语境及句子之间的内部逻辑关系间接地表达出来。而英语常通过各种形式手段连接词语、分句或从句，强调显式连接，各句之间的联系大多是通过词汇语法直接显示出来，使用各种连接词，如 and, but, if, as 等和某些特定短语、分句等^[80]。例如，Zhou 和 Xue^[65]的统计表明，汉语中没有连接词的隐式关系占 82%，而英语占 54.5%。3) 汉英篇章关系分类差异。汉英篇章关系分类体系大不同，乐明^[61]的研究表明使用 RST 标注汉语篇章结构时，有 9.6%的篇章单元间之间需要添加新的关系。4) 汉英篇章中心的确定方法不同，英语中心的确定和主要和关系相关，但汉语篇章中心主要和篇章所要表达的意图相关。例如，同样是因果关系，有时原因比较重要，而有时结果比较重要。

2.1.3 连接依存树

针对汉语篇章结构的一般特点，本文试图结合 RST 和 PDTB 的优点，吸取 RST

的树形结构和篇章单位主次思想，PDTB 的连接词处理方法，参考汉语复句和句群理论等的研究成果，采用一种连接依存树的形式表示汉语的篇章结构⁴。图 2.2 给出了例 2.1.3 的连接依存树表示。由于本文最终在宾州汉语树库（CTB6.0）上采用连接依存树标注汉语篇章树库语料，因此同时给出 CTB6.0 中的一个例子（例 2.1.4）的连接依存树表示（如图 2.3 所示）。

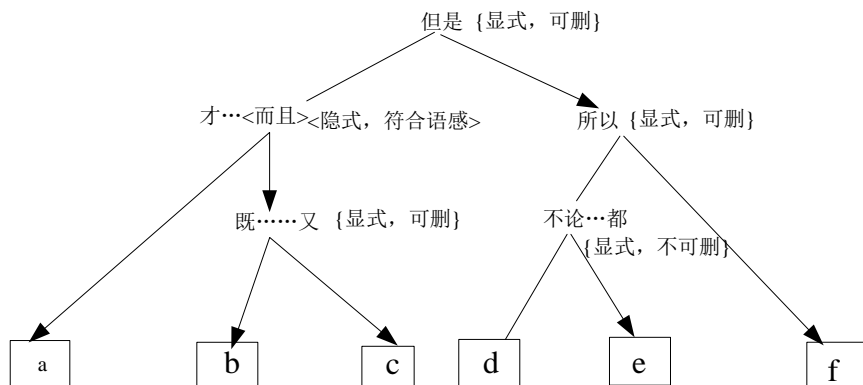


图2.2 例 2.1.3 的篇章结构连接依存树表示

例2.1.3 a 张三才 30 出头，||b<而且>既没有什么学历，|||c 又没有什么新的工作经验。|d 但是不论干什么，|||e 他都非常认真，||f 所以，领导总是把一些重要的任务交给他。

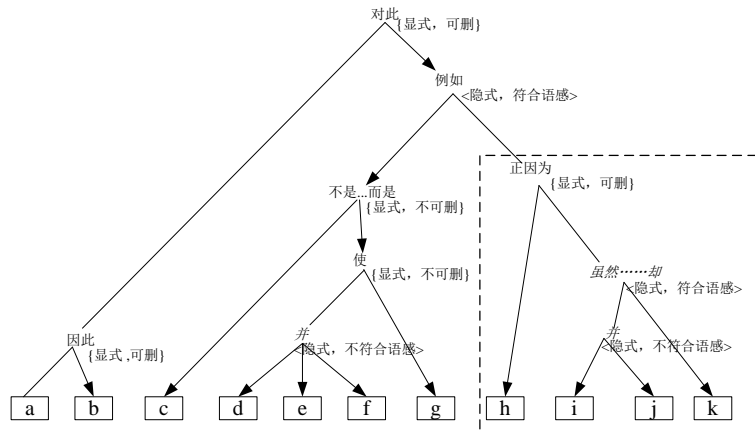


图2.3 例 2.1.4 的连接依存树实例

例2.1.4 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，

⁴篇章结构的研究表明篇章结构适合树形显示，从计算语言学的角度来讲，树结构便于计算分析。篇章单位之间可能存在关系交叉的情况，不过这种情况不多，可以参照依存句法分析，予以合理解决。

||||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 2.2 和图 2.3 中字母所标记的叶子节点表示基本篇章单位 (Elementary Discourse Unit, EDU)，本文称为子句，内部节点表示连接词 (Connective)，内部节点所管辖的子句组合称为篇章单位。各子句之间通过连接词组合后形成高级篇章单位，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 2.2 和图 2.3 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。由于连接词的突出地位，中心又和依存结构相似，本文把这种连接词驱动的生成树称为“连接依存树” (Connective-driven Dependency Tree, CDT)。下面对“连接依存树”进行定义和介绍。

2.2 叶子节点—子句

2.2.1 子句的定义

篇章结构分析中，基本篇章单位的识别是首要任务，对于篇章表示，首先需要确定基本篇章单位。对于什么是基本篇章单位，不同的理论看法并不完全相同：Hobbs 模型^[6]提出，篇章结构由篇章单位和篇章连接关系构成，其中篇章单位可以小到子句，大到篇章本身；Givon^[81]认为从句应该成为篇章的基本单位，Sacks^[82]认为谈话的话轮应该成为篇章的基本单位；Polanyi^[83]坚持篇章应该以自然句为切分单位；Grosz 等^[7]认为篇章的基本单位应该从篇章的上下文中获取，它是由一定的符号所反映的信息载体，能反映事物的单个状态或部分状态；Mann 和 Thompson^[8]的 RST 认为从句应该是篇章的基本单位，无论从句是否存在语法标记或词汇标记，它与 Hobbs 模型有很大的相似性，但相比 Hobbs 模型，RST 更注重句子内部的篇章结构，篇章单元可以小到短语。

自然语言处理领域中，不同的篇章标注语料库对基本篇章单位的定义和标注也有所不同。Carlson 等^[32]综合了 Grosz 等^[12]、Mann 和 Thompson^[9]的理论，在确定基本篇章单位时考虑到词汇、句法、语义和位置等因素。他们的 RSTDT 对基本篇章单位

的规定如下^[33]：充当主语或宾语的从句不属于基本篇章单位；充当主要动词补语的从句不属于基本篇章单位；所有有词汇或句法标记的起状语作用的从句或非谓语动词词组都属于基本篇章单位；定语从句、后置的名词修饰短语或将其它基本篇章单位割裂开的从句或非谓语动词短语为内置篇章单位；有明显篇章标记的短语作为基本篇章单位，如由因为、尽管（in spite of）、根据（according to）等引导的短语属于基本篇章单位。

PDTB 以连接词为篇章级谓词^[15]，篇章单位为谓词论元，一个谓词可以带两个论元。PDTB 的篇章单位包括：简单从句；特殊的非从句（并列动词短语、起名词作用的短语、抽象对象的指代表达形式）；表示连接关系的多个从句或句子（采用最小化原则，即选择最少的能解释连接关系的句子作为论元）。例如“The city’s Campaign Finance Board has refused to pay Mr. Dinkins \$95,142 in matching funds **because** his campaign records are incomplete.”句中，because 前后均为基本篇章单位。

汉语方面，面向自然语言处理的篇章结构分析研究还不多见。乐明^[61-62]把汉语篇章的基本篇章单位定义为小句，形式上小句是由句号、问号、叹号、分号、冒号、破折号、省略号以及段落结束标记所分隔的文字串。陈莉萍^[63]提出的篇章单位应该是以标点符号（如逗号、句号、分号等）分割的句子。综合英语和汉语研究，不难发现汉语篇章结构分析研究首先需要找到一种既能表示汉语本意又便于计算的汉语篇章单位定义方法。

连接依存树的叶子节点为基本篇章单位，本文称为子句（Clause），在从下到上的篇章结构组合中，它是分析的起点；在从上到下的篇章结构分析中，它是分析的终点。由此子句分析相当关键。由于汉语句子的结构和短语的结构没有明显的形式区分，汉语“句”相当难于定义，至今语法学界没有统一明确的看法^[84]。参考主流汉语句法理论，结合可操作性，本文对汉语子句定义如下：

子句含传统单句及复句中的分句。结构上，子句至少包含一个谓语部分，至少表达一个命题；功能上，子句对外不作为其它子句结构的语法成分，子句和子句间发生命题关系；形式上，子句间一定有标点（通常是逗号、分号和句号等）分割。

依据上述子句判断原则，本文用字母标出例 2.1.3 和例 2.1.4 中的子句。为便于区分子句，本文给出例 2.2.1 进行对比说明。例 2.2.1 中，A) 含一个子句，逗号左边非独立命题；B) 含一个子句，有多命题但命题间无标点；C) 含两个子句，每个子句均含命题，

子句间有标点。

例2.2.1 A) a 那个张三, 推开门出去了。

B) a 张三推开门出去了。

C) a 张三推开门, b 出去了。

通过例 2.2.1 可以发现, 标点在于句判断中起非常重要的作用。根据本文定义, 下面对子句判定进行具体说明, 所给例子均出自 CTB6.0, 例后括号指明来源文档编号, 并人工进行了子句划分, 子句前用字母“a、b、……”标示。

2.2.2 子句的判定

2.2.2.1 子句是单句

单句能够表达一个相对完整的意思并且有一个特定的语调, 全句主干只有一个结构中心, 每个结构中心只能有一套句子成分, 即只能有一个主谓结构或非主谓结构。

例2.2.2 a 外商投资企业在改善中国出口商品结构中发挥了显著作用。 (chtb_0001)

例2.2.3 a 北海市的崛起, 是近年来广西壮族自治区对外开放取得卓著成就的重要标志之一。 (chtb_0006)

例 2.2.2 中, 句子从头到尾只在句后有一个标点(句号、问号和感叹号), 这种句子一定是子句。例 2.2.3 中, 虽然句子中间有标点, 但属于只有一套句法成分的单句, 因此也是子句。

2.2.2.2 子句是复句中的分句

复句是由两个或两个以上意义上有密切关系的分句(结构上类似单句而没有完整句调的语法单位)组成的语言单位, 包括简单复句(内部只有一层语义关系)和多重复句(内部包含多层语义关系)。复句中的各个分句之间一般有停顿, 书面上用逗号、分号或冒号表示。

例2.2.4 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。 (chtb_0001)

例2.2.5 a 古老的京杭大运河如今不仅在贯通南北运输方面发挥重要作用, |b 而且带动起一条欣欣向荣的工业走廊, ||c 形成了大运河经济带。 (chtb_0004)

例 2.2.4 中, 虽然子句 a 中间有逗号分割, 但它是复句中的分句, 是一个子句, a 和 b 之间是因果关系, 通过连接词“因此”体现。例 2.2.5 中, bcd 都是分句, b 和 c 是并列关

系, 连接词是隐含的; b 和 c 的组合与 a 是递进关系, 连接词是“不仅……而且……”。

2.2.2.3 标点与子句判定

2.2.1 中子句定义是判定子句的基本原则, 但实际语料中有多种标点, 不同的标点作用也不同, 由于一个子句必然是以标点为边界, 所以标点对子句分割意义重大。文献[85]指出, 汉语的常用标点符号有 16 种, 分点号和标号两大类。点号的作用在于点断, 主要表示说话时的停顿和语气。点号又分为句末点号和句内点号。句末点号用在句末, 有句号、问号、叹号 3 种, 表示句末的停顿, 同时表示句子的语气。句内点号用在句内, 有逗号、顿号、分号、冒号 4 种, 表示句内的各种不同性质的停顿。标号的作用在于标明, 主要标明语句的性质和作用。常用的标号有 9 种: 引号、括号、破折号、省略号、着重号、连接号、间隔号、书名号和专名号。可见, 汉语书面语中和子句边界有关系的是点号, 其中句号、问号、叹号、分号一定表示子句边界; 顿号表示句子内部并列词语之间的停顿, 所以一定不是子句边界; 逗号和冒号可能是, 也可能不是子句边界。CTB6.0 中有可能是子句边界的标点分布如图 2.4 所示, 逗号和句号出现次数最多 (占 86%)。

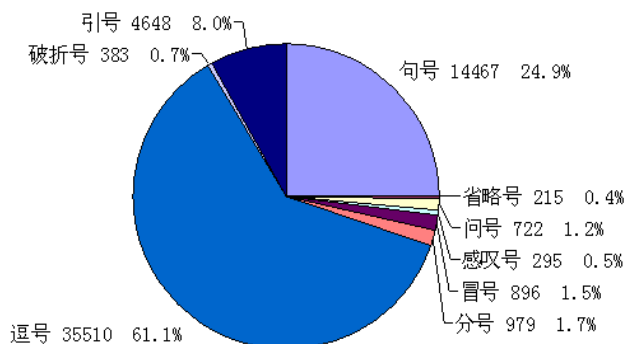


图2.4 CTB6.0 中子句边界标点使用频率比较

每种标点功能不一样, 在进行子句判断时其能否作为子句边界标点的处理方法也不同, 具体如下:

● 逗号

逗号是句内点号的一种, 表示句子或语段内部的一般性停顿。Yang 等^[75]给出的逗号分类方法, 共将逗号的使用方法划分为七类, 首先把逗号的使用方法在总体上分为两大类: 一类是逗号连接的两子句之间存在关系, 即逗号是子句边界; 另一类是两子句之间不存在关系, 即不能标记篇章单位的逗号。两子句之间存在的关系又可以分为并列关系和从属关系。并列关系又可分为三种类型 (SB、IP_COORD 与

VP_COORD), 从属关系也分为三种类型 (ADJ、COMP 与 SBJ)。图 2.5 展示了逗号分类类别。Yang 等的分类主要是从句法树上自动抽取的, 而且其分类的基本单位没有严格的定义, 参考 Yang 的分类方法, 本文将逗号功能分为子句边界类和非子句边界类。

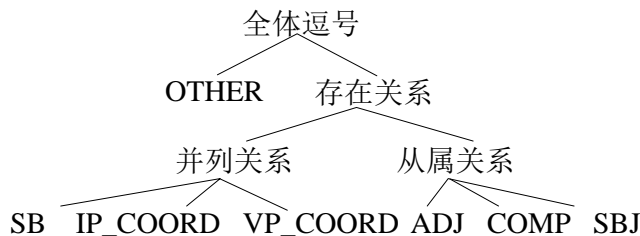


图2.5 Yang 的逗号分类类别

1) SB(Sentence Boundary): 分割句子边界的逗号, 属于子句边界类

SB 类逗号是指在某些语境下, 起到作为句子边界的作用, 这类逗号要求逗号左右的子句都是 IP 结构, 父节点为根节点, 比如在流水句中。例 2.2.6 就是一个流水句, 该句的句法树结构如图 2.6 所示, 例 2.2.6 中两个逗号分别用 P1 和 P2 标示, P1 和 P2 的左右子句均是 IP 结构, P1 和 P2 均属于 SB 类逗号。

例2.2.6 a 陕西省目前批准的外资项目已达二千四百多个 P1, b 协议利用外资额四十多亿美元 P2, c 实际引进外资超过十六亿美元。(chtb_0091)

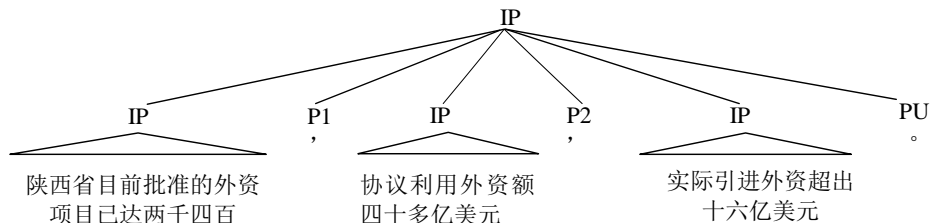


图2.6 例 2.2.6 句法结构

2) IP_COORD (IP Coordination): 分割父节点为非根节点的并列 IP 结构的逗号, 属于子句边界类

该类逗号分割开的子句拥有完整的 IP 结构, 但该逗号的父节点不是根节点, 不能等同于句子边界。图 2.7 展示了例 2.2.7 的句法结构, 该结构通常是长句子的嵌套结构。例 2.2.7 中逗号 P4 和 P5 就属于该分类。

例2.2.7 a 他在会议工作报告中指出 P3, 陆上石油勘探开发遇到一系列世界级难题 P4, b 投资成本日益上升 P5, c 企业改革和产业结构调整任务艰巨。(chtb_0100)

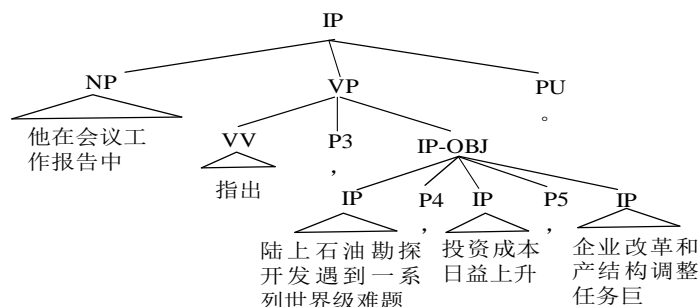


图2.7 例 2.2.7 的句法结构

3) VP_COORD (VP Coordination): 分割并列动宾短语的逗号，属于子句边界类

这一类的逗号与 IP_COORD 类逗号相似，都是分割嵌套结构中的并列结构。VP_COORD 类逗号是分割并列的动宾结构（即 VP 结构）。被该类逗号分割开的子句属于并列的动宾结构，它们共享同一个主语。例 2.2.8 中的逗号 P6 就属于 VP_COORD 类逗号，图 2.8 展示了该例句的句法结构。

例2.2.8 a 中国银行是四大国有商业银行之一 P6, b 也是中国主要的外汇银行。

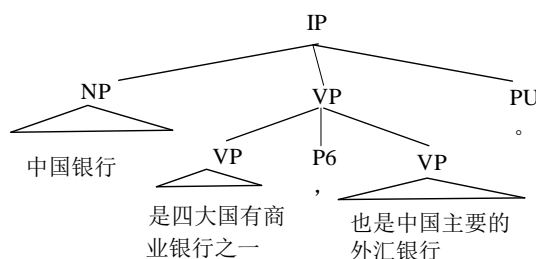


图2.8 例 2.2.8 的句法结构

4) ADJ (Adjunction): 分割附属从句与主句的逗号，本文根据不同情况区别对待

一般情况为非子句边界，但若附属成分是目的状语、原因状语等可以表示独立意义的语句块则认为是子句边界类，否则为非子句边界类。附属从句指在句子中担当某种句子成分的主属结构。虽然从句部分的句子结构是完整的，但它并不能脱离主句部分独立完整的表达意思。附属从句往往是状语从句，通常有条件状语（CND）、原因状语（PRP）、目的状语（PRP）、方式状语（MNR）、伴随状语（ADV，属于其它类型的状语）等。这些附属从句通常情况下位于主句之前，本文也只考察从句出现在主句之前的情况。例 2.2.9 为典型的的目的状语从句，逗号 P7 为 ADJ 类逗号。图 2.9 给出了例 2.2.9 相对应的句法结构。

例2.2.9 a 为了在运行机制上与保护区相配套 P7, b 宁波保护区率先在中国实施了企业依法注册直接登记制的试行一站式管理。

(chtb_0019)

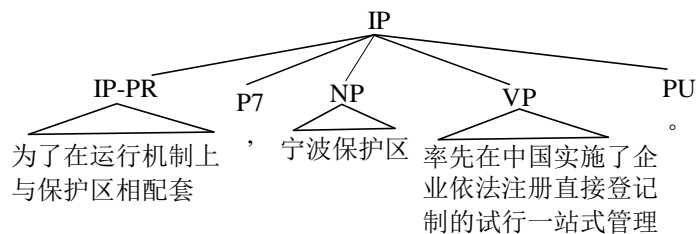


图2.9 例 2.2.9 的句法结构

5) COMP (Complementation): 分割句子谓语与宾语的逗号，属于非子句边界类

汉语中的谓语与宾语之间的联系比较紧密，通常二者之间不出现停顿符号。但对于宾语部分较长的复杂句子，会在谓语之后出现逗号，表示停顿，用于舒缓语气。通常在“表示”、“指出”、“认为”、“介绍”等提示性动词之后都会出现逗号。COMP类就是标示的该类逗号，是一种比较常见的逗号使用方法。例 2.2.10 中的逗号 P8 属于该种情况，图 2.10 给出了该例句的句法结构树。另外，例 2.2.7 中的逗号 P3 也是属于 COMP 类逗号。

例2.2.10 a 钱其琛表示 P8, 我们对香港的前景始终是充满信心的。

(chtb_0058)

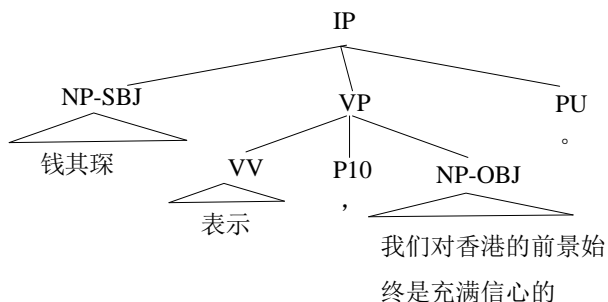


图2.10 例 2.2.10 的句法结构树

6) SBJ (Subject): 分割句子主语和谓语的逗号，本文认为是非子句边界类

SBJ 类表示逗号分割开了句子的主语与动宾结构。在句法结构上表示为逗号的左兄弟节点为 IP-SBJ 或 NP-SBJ 结构，而右兄弟节点为 VP 结构。例 2.2.11 中的逗号 P9 为该类逗号，图 2.11 给出了例 2.2.11 对应的句法结构树。

例2.2.11 a 出口快速增长 P9, 成为推动经济增长的重要力量。

(chtb_0097)

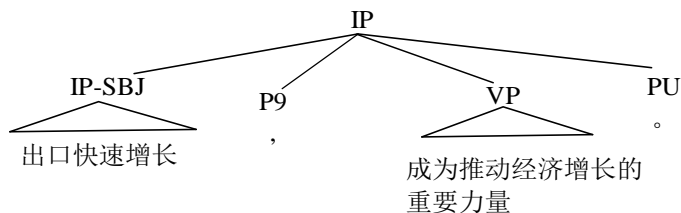


图2.11 例 2.2.11 的句法结构树

7) **Other**: 其它类型逗号, 根据本文子句定义进行划分, 不能采用自动方式进行。

根据上述六种逗号类型的句法特征, 可以很容易地将逗号分类标签提取出来, 不属于上述六种类型的逗号均为 **Other** 类型。

● 句号、问号和叹号

句号、问号和叹号均是句末点号, 句号主要表示句子的陈述语气, 叹号主要表示句子的感叹语气。句号、问号和叹号均为子句分割符, 如例 2.2.12—例 2.2.14。

例2.2.12 a 外商投资企业在改善中国出口商品结构中发挥了显著作用。 (chtb_0002)

例2.2.13 a 雄孔雀时代来临? (chtb_1018)

例2.2.14 a 什么新东西也没有啊! (chtb_1020)

● 分号

分号是句内点号的一种, 分号作为介于句号和逗号之间停顿的标点符号, 用法主要有: 复句内部, 并列分句之间的停顿; 非并列关系 (如选择关系、转折关系等) 的多重复句, 第一层的前后两部分之间; 分行列举的各项之间。在语料中分号出现的比例虽然只有 1.7%, 但语料分析显示, 分号所分割的 99% 以上的句子都是复句中结构完整的分句, 因此, 本文将分号视为子句分割符。

例2.2.15 a 农牧业生产贷款 (包括扶贫贷款) 比上年新增四点三八亿元; b 乡镇企业贷款增幅为百分之六十一一点八三。 (chtb_0007)

例2.2.16 a 贷款将向能源、交通、电力等基础设施产业倾斜, b 尤其以国外大公司在华设立的大中型企业为重点; c 此外, 高技术、高科技、高出口、高利税的企业也将获得中国银行的贷款支持。 (chtb_0007)

例2.2.17 a (甲) 其父或母根据《基本法》第二十四条第二款第一项或第二十四条第二款第二项是香港特别行政区永久性居民;

b (乙) 在儿童出生时, 其父或母是香港或香港特别行政区的永久性居民;

c (丙) 在香港以外出生的儿童是中国籍人士; (chtb_0544)

例 2.2.15 中, 分号前后分别是一个单句, a 和 b 为子句, 分号连接起来的子句之间是并列关系。例 2.2.16 中, a、b 和 c 是子句, a 和 b 是递进关系, a 和 b 的组合和 c 以分号分割, 分号前后都是子句, 表示并列关系。例 2.2.17 用于分项列举的各项之间。

● 冒号

冒号是句内点号的一种, 表示语段中提示下文或总结上文的停顿。冒号在本语料中主要有以下用法: 用于总结性或提示性词语 (如“说”、“报道”等) 之后、用于总结

上文、用在需要说明的词语之后表注释说明、用在列举式或条文式表述中、用在称谓之后。

例2.2.18 a 作者：陈彬

b 出版：商讯文化

c 地址：台北市大理街 132 号 (chtb_1051)

例2.2.19 a 国家统计局预测：全球经济发展将给中国带来很多机遇 (chtb_0016)

例2.2.20 a 江泽民主席在九江视察时说：“九江地处京九中段，b 地理位置很好，c 九江前途无量。” (chtb_0042)

例2.2.21 a 中国应急培训高级跨国经营管理人才的一项重要举措今天在上海实施：b 中国跨国企业经营管理实务高级研修班，今天在上海国际金融学院正式举行开学典礼。 (chtb_0297)

由例 2.2.18—例 2.2.21 可知，冒号用法较多，不能根据标点直接确定是否可以分割子句，需要考虑实际情况进行区分。用于篇章组织和句内的冒号不能作为子句分割符，句间的冒号可以作为子句分割符。

● 引号

引号是标号的一种，标示语段中直接引用的内容或需要特别指出的成分。行文中直接引用的话，用引号标示，如例 2.2.22 中具体艾滋病防治策略；需要着重论述的对象，用引号标示；具有特殊含义的词语，也用引号标示；引号里面还要用引号时，外面一层用双引号，里面一层用单引号。

例2.2.22 a 去年，卫生部和有关部门提出了“预防为主，宣传教育为主，经常性工作为主”的艾滋病防治策略，b 并得到国务院的确认。 (chtb_0244)

对于引号的处理，遵循的原则是若引号位于一个子句内部的（例 2.2.22），不予处理；若引号所引是一个独立的句子（例 2.2.20），将其作为独立句子处理。

● 破折号

破折号是标号的一种，标示语段中某些成分的注释、补充说明或语音、意义的变化。基本用法有标示注释内容或补充说明（也可用括号）、标示插入语（也可用逗号）、标示总结上文或提示下文（也可用冒号）、标示话题的转变、标示声音的延长、标示话语的中断或间隔、标示引出对话、标示事项的列举分承等。CTB6.0 语料中大部分表示后续语句对前面某些词句的解释说明，本文除了表示插入语破折号外，其它情况的破折号均作为基本篇章单位分割符。如例 2.2.23 和例 2.2.24。

例2.2.23 a 这个开发区位于中国著名风景旅游城——杭州市区内, b 是一九九一年国务院批准建设的国家级高新技术产业开发区。(chtb_0011)

例2.2.24 a 不一会, 在一小片开阔地, 导游——一位会操5种欧洲语言的女郎——命令车子停下, b 她让大家趁太阳正露出笑脸向我们致意的时机, 赶紧观赏一下圣岳的容颜。(chtb_0207)

● 省略号

省略号是标号的一种, 标示语段中某些内容的省略及意义的断续等。基本用法有标示引文的省略、标示列举或重复词语的省略、标示语意未尽、标示说话时断断续续等。

例2.2.25 a 这年五月, 九七歌仔戏创作研讨会在厦门召开……(chtb_0836)

例2.2.26 a 处于族群复杂的环境中, 邓相扬对族群间的冲突、融合、文化的消失……等情形非常敏感。(chtb_1011)

对于省略号的处理, 如果其在句尾(如例2.2.25)相当于句末标号, 处理方式参考前面, 如果其在句子中间(如例2.2.26), 则不将其作为子句分割符。

● 其它符号

其它符号不能作为子句边界, 故图2.4没有将其统计进去。圆括号通常用来解释一个特定的词或短语, 不管括号内容中间是否有标点, 均不将其视为基本语篇单位(如例2.2.27)。顿号是中文特有的标点, 表示并列的词或词组之间的停顿, 无论顿号中间的内容多少, 均不将其视为子句分割标点(如例2.2.16、例2.2.26和例2.2.27)。书名号标示书名、篇名、戏剧名、歌曲名、报刊杂志名和法规文件等题名, 属于子句内部符号, 书名号内部即使有其它符号也不必处理。

例2.2.27 a 已经报名采访的记者有1320名(文字601名、摄影171名、电台、电视台400名等)。(chtb_0306)

例2.2.28 a 一九九四年版《经济白皮书: 中国经济形势与展望》, 近日由中国发展出版社出版。(chtb_0268)

2.2.2.4 一些特殊情况

在子句判断时, 通常会有些特殊情况需要单独出来, 主要是言说结构和介词结构。

● 言说结构

“说”类动词引导的结构, 当“说”后引导的内容是一个篇章时, 视其为一般篇章进行结构分析, 并且把“**说”也归到“说”后的第一个子句中。如例2.2.29和例2.2.30的第一个子句a。

例2.2.29 a 他说：“我相信，只要双方共同做出努力，b 两国的友好合作关系就一定会在中法建交公报和今年一月十二日联合公报的原则基础上不断向前发展。”(chtb_0238)

例2.2.30 a 他在会议工作报告中指出，陆上石油勘探开发遇到一系列世界级难题，b 投资成本日益上升，c 企业改革和产业结构调整任务艰巨。(chtb_0100)

● 一些介词引导的结构

当介词引导的结构，在介词删去后可以独立成句，视该结构为子句。如例 2.1.4 中“使这些经济活动一出现就被纳入法制轨道。”删除“使”独立成句，可以视为子句；同样，如例 2.2.31 中的子句 a。

例2.2.31 a 为了在运行机制上与保护区相配套，b 宁波保护区率先在中国实施了企业依法注册直接登记制的试行一站式管理。(chtb_0019)

2.3 内部节点—连接词

在汉语篇章中，篇章关系指同一篇章内部，句子之间或子句之间的语义连接关系，如条件关系、转折关系、因果关系等^[86]，连接词主要指连接不同篇章单位并表示这种语义关系的词语。连接依存树的内部节点为连接词，连接词指具有子句及其以上语法单位连接和关系提示作用的语言单位。判断连接词的主要标准是看其联系的成分是否为子句及其以上语法单位，其能否提示所联系篇章单位间的语义关系。

如例 2.1.4 中的“因此”、“对此”、“不是……而是”、“使”和“正因为”，例 2.3.1 中的“不仅……而且”，例 2.3.2 中的介词“为”，这些对上下句起连接作用的词，都叫连接词。本文所述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，能恰当的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，其范围要广于现代汉语^[87]中的连词。

例2.3.1 a 古老的京杭大运河如今**不仅**在贯通南北运输方面发挥重要作用，b **而且**带动起一条欣欣向荣的工业走廊。(chtb_0013)

例2.3.2 a **为**充分发挥地缘优势，b 进一步加快对外开放步伐，c 呼伦贝尔盟自去年下半年开始，认真总结在对外开放方面存在的问题和不足，d 大力开展了“创造良好开放环境”活动。(chtb_0024)

篇章关系识别是篇章结构分析的基本任务。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，如图 1.1 的做法。本文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篇章关系，如图 2.2 所示，“没……学历”和“没……经验”之间的

关系为“既……又”，进一步二者组合后与“他三十出头”构成“才……而且”关系（对于篇章关系的抽象概括，本文给出一种通用的篇章关系分类体系，但分类体系和连接词相互独立，见2.4节篇章结构关系）。这样，使用连接词既可表示篇章关系，又可避免篇章关系标注中抽象分类与判断的分歧，从而使本文所提的汉语篇章结构表示体系便于实际操作，基于该策略构建的标注语料也便于扩展。本文所称“连接依存树”，与连接词在基于连接依存树的汉语篇章结构表示体系中的独特结构地位和关系表示作用有关。

连接词连接篇章实体单位，在连接依存树中，它一方面通过管辖体现篇章的层次结构，另一方面通过其语义提示篇章的逻辑语义关系。连接词的判定是连接依存树表示体系的基本任务之一，充分认识连接词的特点有利于准确地判定连接词。下面分别从连接词的特点、连接词添加删除及其逻辑关系对连接词进行详细介绍。

2.3.1 连接词的特点

作为篇章中的连接词，最重要的作用是连接前后子句或句子等篇章单位。如例2.3.1句中的“不仅……而且”，例2.3.2句中的“为”就很好的起到了连接逗号前后子句的作用。此外，连接词还具有其它性质，下面从连接词的形式、分布、词性和句法等方面进行说明。

2.3.1.1 连接词的形式

在汉语篇章结构中，连接词的形式并不是单一的。根据其形式的不同可以分为独用连接词、关联词、合用连接词和其它类型。

独用连接词指在汉语篇章结构中，连接子句或句子之间的词是单独的词语，如“并且”、“而且”、“所以”等。关联词指在汉语篇章结构中，连接子句或句子之间的连接词是成对儿出现的词语，如“既……又”、“不但……而且”等。合用连接词是指几个词组成短语，连接前后子句，表示同一篇章关系的词，如“同时还”，具体见例2.3.3。

例2.3.3

A) a 据介绍，国家开发银行成立后，将专业银行的这部分用于基建和技改项目的信贷基金以金融债券的形式转到国家开发银行名下，b 所以在固定资产投资总量上并没有发生变化。

（因果关系）

(chtb_0218)

B) a 他说，由于人民币资本项目下的可兑换本来就没有时间表，b 所以不存在因东南亚金融危机而延长这一过程的问题。（因果关系）

(chtb_0123)

C) a 专家们表示, 福建发展高新技术产业, **不仅能**加速本地产业的升级换代和国际化, **b 而且**最终有可能使这一地区成为海峡两岸科技、经贸合作的最佳地带。(递进关系)

(chtb_0034)

D) a 目前, 已有十二万多家外商投资企业在中国开业, **b 而且**这些已开业的外商投资企业绝大部分生产经营状况较好。(递进关系)

(chtb_0006)

E) a 今后还将适当增加外国银行和保险公司在中国的分支机构, **b 同时还**准备扩大外国银行办理人民币业务的试点。(并列关系)

(chtb_0123)

F) a 他指出, 美国国会每年就这个问题进行辩论实际上只有对美国自身不利, | **b 影响**美国商人的对华投资信心, || **c 从而也**影响到美国人的就业机会。(因果关系)

(chtb_0146)

G) a 入冬以来, 已有四百余只丹顶鹤陆续飞抵江苏盐城沿海滩涂越冬。| **b 然而由于**少数人法制观念淡薄, || **c 当地已**连续发生数起毒杀丹顶鹤事件。(然而—转折关系; 由于—因果关系)

(chtb_0654)

分析例 2.3.3, A)中连接词“所以”在这里是独用连接词, 表示因果关系, 但“所以”很多情况下可以与其它词联用表示因果关系, 如和 B)中的“由于”联用表示因果关系。B)连接词“由于……所以……”和 C)中的连接词“不仅……而且……”都是成对相互关联出现的, 为关联连接词。但是关联词也可以独立应用到子句中, 单独表示子句间的连接关系, 如 D)中“而且”和 A)中“所以”。E)中的“同时还”是合用连接词, “同时”和“还”两个连接词合并起来表示并列关系。F)中的连接词“从而也”是表示因果关系的“从而”和表示“并列关系”的“也”合用, 在所连接的子句中表示因果关系。G)中的“然而由于”虽然相邻, 与合用连接词类似, 但“然后”和“由于”分别有不同的管辖范围, 表示不同的关系, “然而”表示子句 a 和 bc 组合是转折关系, “由于”表示子句 b 和子句 c 是因果关系。

2.3.1.2 连接词的分布

从分布上看, 连接词可以出现在连接项的开头、中间和结尾。例 2.3.4 中“只有”、“虽然”、“如果”位于句首。例 2.3.4 中“才”、“但”, 例 2.3.3 中“所以”、“而且”位于句中, 子句开头。例 2.3.4 中“外”位于子句尾部。

例2.3.4

A) a **只有**实现持久的和平, **b 才**有可能实现持续发展。(chtb_0052)

B) a **虽然**一九八三年这个村也开始实行家庭联产承包责任制, **b 但**整个八十年代, 大寨都处在沉默之中。(chtb_0644)

C) a **如果**100 条新闻中有一条是假的, **b 读者**对另外99 条也会产生怀疑。(chtb_0196)

D) a 除建立合资企业外, b 通用**还**通过飞机和集装箱租赁向中国投入十亿多美元。

(chtb_0028)

2.3.1.3 连接词的词性

从语法性质上看,连接词不限于传统连词,只要是起子句及其以上语法单位连接和关系提示作用的语言单位均为连接词,如例 2.1.4 中的连接词有连词(“因此”)、介词(“使”)、动词短语(“不是……而是”)、介词短语(“对此”)等。

在汉语篇章结构中,连接词有连词、介词、副词等诸多语法类型。其中,占绝大多数的还是连词。连词起连接词、子句和句子等的作用,表示并列、选择、递进、转折、条件、因果等关系。例如:和、跟、同、与、及、或;而、而且、并、并且、或者;不但、不仅、虽然、然而、如果、因为、所以。介词依附在实词或短语前面共同构成“介词短语”,通常用于修饰、补充谓词性词语。介词常常充当逻辑成分,表明与动作、性状有关的原因、目的、方式和处所等。例如:因为、由于、为、为了、除了、对于、关于。在现代汉语中,副词是修饰限定动词和形容词,表示程度、范围、时间等意义的词。如:更、更加、还、还是、只、仅仅、尤其等。在汉语篇章结构中,利用副词来表示子句之间的递进、顺承等关系。如“尤其”作为连接词可以表示子句间的递进关系;“再”可以表示顺承关系等。在上述词类中,连词在连接词中所占比例最高,其次是副词,介词比例最低。

2.3.1.4 连接词的句法特性

作为连接词,在篇章中一般只起连接作用,不充当句法成分。如例 2.3.1 中的“不仅……而且”,词语本身在句中对前后子句并不起修饰作用,只起到连接前后子句的作用。但也有既充当句法成分,也起连接作用的连接词,如例 2.1.4 中“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,而是借鉴发达国家和深圳等特区的经验教训……”的“不是”,在子句中既起否定作用,也对子句起连接作用。根据词的意义和所起的作用,对该类词做出判断也不难,此类连接词并不多。

2.3.1.5 连接词的逻辑语义关系

一般连接词没有实实在在的意义,不像名词、动词、形容词具有一定的内涵,如“但是”、“不但……而且”、“因为……所以”等连接词,起到连接前后子句的作用,词语本身并无实际意义。但连接词本身可以表示一定的逻辑语义关系,如“但是”表

示转折关系,“因为……所以”表示因果关系,“不仅……而且”表示递进关系等等。在表示篇章结构逻辑语义关系时,连接词和关系类别并不是一一对应的,一个关系类别可以有很多连接词,如因果关系中可以包含“因为……所以”、“因此”、“因而”、“于是”等连接词。但一个连接词也可以表示多种关系类别,如连接词“而”可以表示并列关系,也可以表示转折关系。关于连接词的逻辑语义关系,本文2.4.2节中有详细的论述。

2.3.1.6 其它连接词

汉语篇章中的有些连接词可以表示多个关系,如连接词“并”既可以表示并列关系又可以表示顺承关系。在实际判断过程中,通过连接词对句意关系的判断不明确,会有同时可以是甲关系和乙关系的情况,一个连接词同时表示多个关系,这类连接词称之为特殊连接词。这类连接词有“而”(例2.3.5列出“而”表示转折关系、并列关系等的实例)、“尽管……但”(转折关系、让步关系)、“无论……还是”(让步关系、并列关系)、“只要……就”(假设关系、条件关系)等。

例2.3.5

A) a 外商投资企业的出口商品仍以轻纺产品为主, b 其中, 出口额最大的是服装, c 去年为七十六点八亿美元。d 而进口商品则以机械设备和工业原材料为主。(转折关系)
(chtb_0002)

B) a 推动经济增长的主要因素是亚洲地区经济发展依然强劲有力, b 全地区经济增长速度将达到百分之七点九, c 而中国增长速度可高达百分之九点七。(递进关系) (chtb_0016)

C) a 出口得到恢复, b 而进口将会减少。(并列关系) (chtb_0122)

D) a 预计今年发展中国家经济增长百分之五点九, b 而明年轻仅增长百分之四点九, c 比10月份的预计低了一点三个百分点。(顺承关系) (chtb_0067)

E) a 一九九七年, 中国的人均储蓄存款超过三千元, b 而二十年前每人每年平均才存二十元钱。(对比关系) (chtb_0651)

2.3.2 隐式连接词的添加

根据连接词是否在篇章中出现,可以将连接词分为显式连接词和隐式连接词两类,连接词的添加是连接依存树表示体系中需要处理的主要问题。统计表明,汉语隐式关系所占比例较大(Zhou 等^[65]的统计表明,隐式关系占82%)。本文将连接词作为关系表示和篇章树构建的纽带,因此,隐式连接词的添加非常关键。虽然汉语连接词缺

省较多,但连接词的添加还是很有必要和可操作的,因为汉语划分关系类别的原则是“从关系出发,用标志控制”^[71],关系指子句之间或句子之间的相互关系,标志指联结子句或句子标示相互关系的连接词。关系属于隐含的语义范畴,理解起来有灵活性,而连接词则是一种客观存在的形式实体,因而可以成为客观标准。

例 2.3.6 中连接词“仅”、“同时”、“并”、“因此”为篇章中出现的连接词,是显式连接词;连接词“但是”、“如果”、“那么”是添加上的连接词,为隐式连接词,用符号“<>”标示。

例2.3.6

A) a 现在,全国已有一千一百九十四个县(市)对外开放,|||b{仅}一类对外开放口岸便达二百二十二个。|c(同时),“八五”时期的对外开放在深度上也创下了历史之最,||d 过去中国的对外开放主要是以商品贸易、技术引进及合资合作为主,|||e<但是>如今已开始向引进服务、引进现代资本运作方式等高层次迈进,||||f(并)开始向海外输出资本,||||甚至开始参与国际金融运作。(chtb_0032)

B) a <如果>东盟同中日韩加强合作,共谋发展,||b<那么>将是举足轻重的力量。|c(因此),东亚首脑非正式会晤的举行对推动各国之间的合作,维护地区和全球安全,加速世界向多极化发展,具有重要意义。(chtb_0052)

隐式连接词的添加不是随意的,添加上的连接词需要正确反映原始关系,下面详细介绍隐式连接词添加的依据。

2.3.2.1 添加连接词的依据

添加连接词,首先要保证标点符号前后部分为子句,子句之间是隐式篇章关系,缺少连接词。不是所有的隐式关系都可以添加上连接词,如果前后关系无法确定,则无法添加连接词。确定所添加的连接词,首先要根据前后句意判断子句间的关系类别,只有关系确定了才容易找到相应的连接词,即使有时是凭借语感添加的连接词,潜意识中子句间的关系已经在添加前确定。添加连接词,必须按照先前划分好的层次,逐层添加,不能按子句从前至后添加。因为每一层可能不止一个子句,层与层之间的关系不同,关系是从高到低的,所添加的连接词管辖范围也有大有小,只有按照层次来添加,才能体现出子句间清晰的脉络关系。下面以例 2.3.7 为例说明连接词的添加依据。

例2.3.7

A) a “八五”期间,广东电子工业优化地区布局。(chtb_0033)

B) a 他认为这是一次具有重要意义的会议,|b<因此>中国政府予以高度重视,||c<然后,

并>将派代表团出席会议。(chtb_0281)

C) a “中国国家气象局购买美国克雷公司的大型计算机, b*<但是>克雷公司只卖给我们两台处理器。(chtb_0040)

D) a 随着中国经济的不断发展和对外开放的不断深入, 外商来华投资热情很高, b<由此, 并且>投资项目和金额增长十分迅速。(chtb_0006)

E) a 今年已新签海外工程承包及劳务合同五百八十六份, b<>合同金额二点六五亿美元。(chtb_0059)

F) a 国际货币基金组织 21 日在此间发表一份临时评估报告, b< >再次调低了它对今明年全球经济增长速度的预测。(chtb_0067)

例 2.3.7 中, A)逗号后面部分虽然是一个完整的句子, 但前面部分不是子句, 缺乏独立性, 前面部分需要依靠后面的子句为支撑, 整体才能表达完整的句意, 因此中间都不能添加连接词。B)中 a 子句、b 子句和 c 子句句意完整, 符合本文子句判断原则, a 和 bc 之间是因果关系, 可以添加连接词“因此”, 也比较符合汉语语感。C)中 a 和 b 之间是转折关系, 可以添加连接词“但是”。

如果对子句之间的关系既可以理解为甲关系, 也可以理解为乙关系, 则所能添加的连接词可以不止一个。对于有多种关系, 可以添加多个连接词的情况, 添加的方法是按照关系强弱依次添加连接词, 最多可以添加 3 个连接词。如例 2.3.7, B)中子句 b 和 c 可以理解为顺承关系, 添加连接词“然后”, 也可以理解为并列关系, 添加连接词“并”, 按照语义关系强弱添加“然后, 并”; D)前后可以理解为因果关系, 添加连接词“由此”, 也可以理解为并列关系, 添加连接词“并且”。

并不是所有的子句之间都可以添加连接词, 很多情况找不到合适的连接词进行添加, 此种情况并不强行添加连接词, 所填连接词可以为空。如 E)中有 a 和 b 两个子句, 有隐式的并列关系, 但找不到适合的连接词进行添加, 故可以不添加连接词, 只是标注其语义关系。F)中子句 a 和子句 b 是解说关系, 但也不好添加合适的连接词, 因此也不强行添加。所以有些子句之间或句子之间根本就不用添加连接词, 本身前后单位句意明晰、通顺流畅, 添加反而使篇章显得啰嗦、累赘。

其实, 判断子句之间的逻辑关系, 确定子句之间的关系类和添加连接词是相辅相成、互相影响的。如例 2.3.7 中的 D)句子, 可以理解为先判断其为转折关系, 由此添加的连接词“但是”, 这是正常的判断思维; 也可以说是从逻辑上感觉添加“但是”合适, 然后转折关系自然而然就体现出来。其实二者之间的区别并不明显。但无论是哪一种模式, 其判断的过程都是极快的, 基于对母语的熟悉和敏感度, 思维在大脑中

瞬间就已完成。

2.3.2.2 连接词添加的位置

根据 2.3.1 中连接词的特点可知,连接词的位置可以是子句的句首、句中,并且位于句首的居多。连接词添加的位置同样遵循连接词出现的位置特点,如例 2.3.8。

例2.3.8

A) a 中国进出口银行聘请日本野村证券公司作顾问, b<然后>向日本著名的评级机构日本公社债研究所提出正式评级申请。 (chtb_0010)

B) a<虽然>成交药品一亿多元, b<但是>没有发现一例回扣。 (chtb_0001)

例 2.3.8 中, A) 句单个连接词“然后”位于 b 子句的句首; C) 句中“虽然……但是”成对儿出现,分别位于 a 子句和 b 子句的句首。

连接词所能添加的位置和连接词本身的活泼性有关。活泼性指连接词所能出现的位置。如果连接词所能出现的的位置较固定,只能位于句首或句中,则其活泼性就较差,较为稳定,可称该类连接词为稳定型连接词。相反,那些既可以位于句首又可以位于句中的连接词活泼性就较强,就可称之为活泼型连接词。

不论所添加的连接词是在句首还是句中,都要确保句子语意连贯、没有歧义、语感流畅通顺。

2.3.2.3 其它情况

连接词的添加是在子句间,但并不是所有的子句之间都可以添加连接词。以下几种情况不可以添加连接词。

● 子句内部

由 2.2.2 可知,连接词的添加只能在子句与子句间,子句内部还包括子句内嵌的句子,不做切分。如例 2.3.9。

例2.3.9

A) a 据不完全统计,广西仅与西南三省一区(四川、贵州、云南省和西藏自治区)实施了一千多个协作项目。 (chtb_0008)

B) a(随着)崇明海关办事处的设立,崇明县内的单位足不出岛就可以办理一切海关手续, b <而且>这对进一步改善崇明县的投资环境,加快吸引外资,方便快捷地办理海关手续,把崇明建设成对外高度开放的大型贸易港口,带动出口加工、航运中转等外向型经济的发展,将起到积极的作用。 (chtb_0007)

例 2.3.9, A) 中逗号后面就不能添加连接词,前后共同构成一个完整子句的整体。

B)中 b 子句内嵌的部分内部结构复杂,暂不做处理。

● 特殊的标点

层次的划分,子句的切分只到逗号,往下不再作处理。连接词的添加是以层次为基础,低于逗号的也不再添加连接词。特殊的标点主要有顿号、引号,如例 2.3.10。
例2.3.10

A) a 使中国在吸引外资、引进技术、拓展市场及发展高科技等领域也面临激烈的国际竞争。(chtb_0016)

B) a 宋健说:“如今,中国已能生产上万门数字电话程控交换机, b (而且)这种交换机的总设计师只有二十八岁, c 武汉大学毕业的。”(chtb_0040)

例 2.3.10 中,虽然 A)句中“吸引外资”、“引进技术”、“拓展市场”都是完整的句意表述,但由于其连接标点是顿号,不是子句,因此无需添加连接词。B)句“宋健说:”不添加连接词,单“宋健说”并不是一个独立的子句,其说的内容在后面,前后才是一个完整的整体。

● 逻辑关系不明确

例 2.3.11 中段落虽然能够划分层次,但前后子句关系不明确,在句意上联系不大,不容易添加连接词。

例2.3.11 a 到去年底,全区各项存款余额达七十一亿六千三百万元, b 比上年同期增长百分之四十一亿七千八百, c 其中,城乡居民储蓄存款为十九亿三千七百万元, d 比上年同期增长百分之四十八点二。(关系不确定) e “八五”期间各项存款比“七五”末净增五十亿元, f 年平均增长百分之二十七点四九。(chtb_0005)

● 特殊句式

有些句式从表面上看符合子句划分的要求,但都不是单独的子句,句意也不完整。这些情况不做层次切分,因此也不添加连接词。主要有:

(1) “据悉、报道、了解、有关部门说、介绍……等等之类的”还有与功能作用相似的“分析人士指出、尤为值得一提的是”等。

(2) 表示时间的,如“在……期间,几年来,去年,1995年”等。

(3) 介词短语表示的句子,如“在……看来、除……外、自/从……以来/开始、包括……在内、……后、在……中、随着……”等。

(4) 句子后面是以“的”字结尾的短语,可以理解为是伴随状语或“的”字的名词性短语,如例:“实行拍卖的,可减免有关税收”,“出售给法人和社会自然人接受职工安置的,买断国有职工工龄的费用”。

(5) 表示“某某人说”不单独作为一个子句，与后面说的内容临近那一层归为一层。

(6) 表示“前者是……”即有某种评价关系的，“是”前面不当做一个子句，不再划分。如“产品、项目水平高，是该区的重要特点”。

2.3.3 显式连接词的删除

连接词的删除是针对篇章中的显式关系来说，因为显式关系中子句已带有连接词，需要判断其能否删除。通常连接词删除以后，不影响其所在子句的独立性。个别充当了主要句法成分（如谓语）的连接词删除以后，可能影响其所在子句的独立性，这样的连接词一般是成对出现的，如例 2.3.12 A) 中的“不是……而是”。

如果连接词删除后，句意没有发生变化，子句依旧独立成句，流畅通顺，不影响句子意思的完整表达，则可以删除，可删除连接词用“()”标示。否则不能删除，用“{ }”标示，如例 2.3.12。

例2.3.12

A) a 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，
b{而是}借鉴发达国家和深圳等特区的经验教训。 (chtb_0001)

B) a 古老的京杭大运河如今(不仅)在贯通南北运输方面发挥重要作用，||b{而且}带动起一条欣欣向荣的工业走廊，|c<由此>形成了大运河经济带。 (chtb_0013)

例 2.3.12 A) 句中“不是……而是”表示转折，都不能删，“不是”不仅是连接词，而且参与句子成分，在句意上表否定意思，如果删去，句子意思就会反转。“而是”如果删去，句子缺少过渡成分。B) 中“不仅……而且”也是成对出现的连接词，表并列关系。“不仅”可以删去，不影响句意表达，“而且”不能删去，因为“而且”删除后子句 a 和 b 的递进关系会消失。

首先，同连接词的添加一样，删除连接词的前提必须是起子句之间连接关系的连接词，如果是子句内部，只在句中起连接前后部分的作用，则不是篇章连接词，也不能删除。本文研究的篇章结构，最基本的单位就是子句，层次的判断、连接词的添加和删除、篇章主次单位的判断等，最小的单位都是子句。如例 2.3.13 中的“和”是一个连词，但只连接该子句内部，其连接范围不在子句与子句间，这种情况“和”不属于篇章连接词，因此不考虑其能否删除的情况。

例2.3.13 a 随着俄罗斯国内对工业品需求向中高档方向发展~~和~~国内经济形势的逐步稳定，
天津市众多的“三资企业”产品正在积极寻求进入俄方市场。 (chtb_0035)

其次是根据句意。其实连接词的删除和连接词的添加共同之处就是依据句意，看原来句意是否完整，子句间关系是否改变，但句意是个抽象的概念，不好直观表述出来。

2.4 篇章结构关系

篇章关系分析是篇章结构分析的基本任务。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，如图 1.1。连接依存树的表示形式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篇章关系。如图 2.2，“没……学历”与“没……经验”之间的关系为“既……又”，进一步二者组合后与“他三十出头”构成“才……而且”关系（对于篇章关系的抽象概括，本表示体系将建立另外的篇章关系对应模块单独处理，见第 2.4.2 节篇章关系）。如此，连接词本身既可表示篇章关系，又可避免篇章关系判断中抽象分类与判断的分歧，从而可获得高度一致的汉语篇章结构表示。

2.4.1 篇章结构层次化及判定

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内部节点上。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次结构的确定。篇章的层次结构本质上反映的是篇章单位间语义关系的紧密程度，表现为子句之间组合的先后顺序。如例 2.1.4 中第 3 句话的子句 *i* 和子句 *j* 关系比较紧密，首先组合；子句 *ij* 的组合又和子句 *k* 比较紧密，进而组合；*h* 和 *ijk* 的组合比较紧密，再组合。依次根据紧密程度，可得到图 2.3 虚框中第 3 句话的结构子树。

由图 2.2 和图 2.3 可知，篇章结构是一种层次化的树形结构，其中基本篇章单位居于最低节点，连接词居于不同层级的高层节点。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核心任务，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才能确定。

篇章结构的构建可采用从下到上的组合策略，也可采用从上到下的切分策略。在篇章结构层次分析中我们采用了从上到下的切分策略。这种策略是在结构分析中自然

选择的,它可能更符合人的普通结构认知方式,也更符合汉语的结构认知特点。这种策略可以使分析者尽量避开子句的定义与切分困扰,从而获得一个比较可靠的宏观篇章结构。

篇章的层次结构本质上反映的是篇章单位间语义关系的紧密程度。篇章单位的关系越远,其结构的层级地位越高;关系越近,其结构的层级地位越低。篇章层次结构判定的根本标准就是篇章单位间语义关系的紧密程度。通常可以借助标点、词汇、句法等形式特征对篇章单位间语义关系的远近及结构层次进行判定。篇章层次例子中用“|”的个数表示所在的层次。具体判定方法有:

第一,利用标点符号的层级地位进行篇章结构分析。首先,可根据句末点号(如句号、问号等)和句内点号(如逗号、分号)的差别区分出句间分析和句内分析。从上到下的分析中,通常先进行句间分析,然后再进行句内分析。其次,可根据句内标点的层级地位(一般是分号高于逗号),进行复句内子句的层级分析。如例 2.4.1A)段,可首先在句号间进行句子的层次切分。进而,在 A)第二句内,可根据分号与逗号层次判定子句间的结构。不过,标点符号的层次作用只是辅助性的,根本上还需凭借句间关系来分析篇章结构。这是因为,一方面同样的标点符号(如逗号、句号)所切割的篇章单位地位不一定相同;另一方面,在实际篇章中,标点符号也不一定完全按照既定的层级顺序使用,如例 2.4.1B)中,分号间的层次就低于某些逗号间的层次。

第二,利用篇章单位的词汇关系进行篇章结构分析。语段间有相同、相近或相对等关系的词语,意味着相关语段的关系可能比较密切。如例 2.4.1A)中,第 2、3 和 4 句的首子句中“建立”、“基金”、“资金”和“基金和资金”等词汇间的相同、相近或包含关系,对于判定它们之间的关系紧密程度及层次是有帮助的。

第三,利用篇章句法结构的异同进行篇章结构分析。语段的句法结构相同、相近,意味着语段关系可能比较密切。如例 2.4.1 A)中第 2、3 句首,子句“建立……”动宾句法结构的运用,例 2.4.1 B)中 ab 子句“举行了”和“对”句法结构的运用,均对判定相关句子间的层次结构有所帮助。

第四,利用连接词的管辖范围进行篇章结构分析。如例 1.2 中,根据“既……又”、“才……而且”中配合使用词语的所在位置,可以较清晰地判定子句 a、b 和 c 之间的层次关系。

例2.4.1

A) a 广东省各级政府近几年不断加强对科技的投入, || b 初步建立起多层次、多渠道的

科技投入新体系。|c 广东省建立了自然科学基金, |||d 每年投入在一亿元以上; ||e 省级用于新产品开发等科技三项经费每年以百分之十的速度增长, |||f 高于全省财政收入的增长速度。|g 近年来, 该省{又}建立了成果转化科技风险资金、科技创业投资资金和高新技术产业发展资金, |||h 一些市、县{还}设立科技发展基金等。||i 这些基金和资金的投入, 有力地支持了省重点实验室和各工程技术研究开发中心的建设, |||g<并>促进了科技成果的产业化。
(chtb_0084)

B) a 今年以来, 联合国开发计划署在中国吉林, 举行了图们江地区国际旅游会议, ||b<以此>对中国长白山冬季冰雪旅游等项目积极支持; |c(同时){还}出资在这里培训国际经贸人才。
(chtb_0049)

这里需要特别说明篇章结构与连接词层级的关系。如图 2.2, 表面上篇章结构表现为连接词的不同层级地位, 但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。所以篇章结构分析的根本任务是分析篇章单位的组合层级。连接词的层级分析虽然可以促进篇章结构分析, 但并不意味着篇章结构分析必然依赖于连接词的层级判定, 甚至可以由连接词的层级或管辖分析代替。在结构分析过程中, 分析者自然地选择的分析流程是先进行结构分析, 然后才把相应的连接词挂靠到其所归属的层级上。该操作流程反映的本质是, 通常连接词的层级判定依赖于篇章单位的层级判定, 而不是相反。

还有一点值得指出, 正因为篇章结构的构建不依赖于连接词的层级或管辖确定, 所以总能构造出篇章的完全结构树。在这一点上, 本文和同样采用了基于连接词思想的 PDTB 有着根本不同, PDTB 的篇章结构完全依赖于连接词的管辖而定, 篇章结构是隐含的; 又因为其对隐式关系分析不完全, 所以一般情况下在 PDTB 中看不到一棵完整的篇章结构树, 当然 PDTB 可能也没有以构造一棵完整的篇章结构树为目标。所以虽然我们和 PDTB 一样利用了连接词的管辖表示篇章的层级地位, 但二者的宗旨和结果有着质的不同。

2.4.2 篇章关系类别及判定

通常情况下, 一个连接词表示一种篇章关系, 但也存在个别连接词歧义或多义的情况。比如“于是”可能表示顺承关系, 也可能表示因果关系。而在一个具体篇章中, 可能存在“于是”仅表示其中一种关系的情况, 也可能存在“于是”同时兼表两种关系的情况。对于一个连接词对应多种篇章关系的情况, 可以按照一般的词义排歧方法进行处理, 也可以对连接词标注语义进行区分。值得指出, 由于连接词的层级管辖判

定和篇章关系表示判定被作为两个相对独立的任务，无论连接词是否多义，表何种意义等，均不会影响篇章结构本身的标注。

图 2.12 给出了我们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，连接词可以分为四大类：因果类、转折类、并列类和解说类。每一类内部又可以细分为不同的关系类型。因果类包含因果关系、推断关系、假设关系、目的关系、条件关系、背景关系六种关系；转折类包含转折关系和让步关系两种关系；并列类包含并列关系、顺承关系、递进关系、选择关系和对比关系五种关系；解说类包含解说关系、总分关系、例证关系和评价关系四种关系，共 17 小类关系。每一种关系类型都包含若干连接词，有的连接词可属于不同的关系类型。

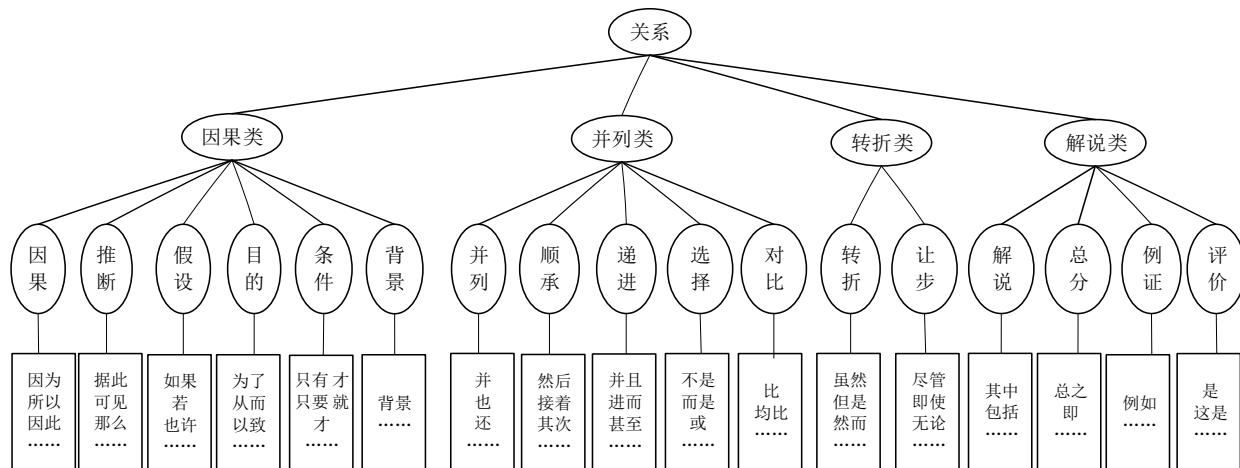


图2.12 基于连接词的关系分类

从级别上说，因果类等 4 大类是一级类别，因果关系、并列关系、转折关系等 17 类关系是二级类别。4 大类只是为一个大群划出一个范围，因果、并列、转折等 17 类关系是具体的关系类型，二级类别下可以根据需要再划分三级类别。原则上，篇章关系的层级和类别都不是封闭的，均可根据具体应用目的或情况进行增加或调整。由于将篇章结构和篇章关系分开，使得基于连接依存树的汉语篇章结构表示体系具有广泛而灵活的适应性。下面详细介绍本文给出的篇章关系类别。

2.4.2.1 篇章关系类别

四大类共有十七种具体的篇章关系，理解篇章关系定义才能更好的分析篇章。小类的划分以子句间的关系为主要依据，除了语义方面，另一个重要原则是注重形式，以联合子句间形式标记即关联词为辅助手段加以分类。以下是对于每一种篇章关系的

定义及相应的释例，用“[]”表示关系类别。

● 并列类

并列类反映各种各样的并列聚合，各子句之间地位是平等的，可以从子句组成成分的数量、次序、意义关系方面进行定义（参考汉语复句^[71]、汉语句群^[59]和修辞结构理论^[9]）。

并列关系，并列项是叙述相关的几件事情或同一事物的几个方面，相关是在意义上并存、平行或对立。并列项可以是词、短语，也可以是句子。并列项可以互换位置，并列项逻辑上不相互依赖，如例 2.4.2 中，各子句间是并列关系。

例2.4.2 a 重庆位于四川盆地东南部，[[并列]b 东西宽二百零八公里，[[并列]c 南北长二百二十公里，[[并列]d 市区座落于长江、嘉陵江交汇处的山丘坡地上，[[并列]e 市域面积二万三千一百一十四平方公里。(chtb_0144)

顺承关系，表示前后句子或子句按时间、空间或逻辑事理上的顺序表达连续的动作或相关的情况。它们之间有先后相承的关系，各子句间次序不能随意改变，如例 2.4.3。

例2.4.3 a 一九六四年，金川公司产出第一批电解镍。|[顺承]b 从此以后，逐步改变了中国镍、钴及铂族金属长期依赖进口的局面。|[顺承]c 如今，这里已成为中国最大的镍钴生产基地和铂族金属提炼中心，||| [并列]d<并且> 镍和铂族金属产量分别占全国的百分之八十八和百分之九十以上，|| [因果]e<因此>被誉为中国的“镍都”。(chtb_0068)

递进关系，后项的意义比前项的意义更进一层，一般由少到多，由小到大，由轻到重，由浅到深，由易到难。强调程度的加深，如例 2.4.4。

例2.4.4 a 进入 12 月以后，欧佩克原油价格跌破了每桶 10 美元大关，|[递进]b(而且)仍在一路下滑。(chtb_0705)

选择关系，选择项分别表述两种或几种可能的情况，从中选择，如例 2.4.5；或选定其中一种，舍弃另一种，如例 2.4.6。

例2.4.5 a 还有一些国际投机资本，混入经常项目汇入境内，|[选择]b{或}入境内股市炒作，|[选择]c {或}在中国境内购买假报关单购汇套取汇差。(chtb_0137)

例2.4.6 正规的传销公司不会要求加入者一定要一次买足大量的货，|[选择]{或}是缴交昂贵的入会费。(chtb_0590)

对比关系，是两种事物或一个事物的两个方面相互比较，如例 2.4.7。

例2.4.7 a 据《中国青年报》报道，深圳姑娘嫁往香港最多的一九八八年是一千零六十九

人,|[对比]b{而}今年一至十一月份仅有二百九十九人。(chtb_0717)

● 转折类

反映各种转折聚合关系,转折类子句组成成分间地位不平等,强调意义的翻转,包括转折关系和让步关系。

转折关系,前后项以某种客观存在的事实为前提,两项意思相反或相对,即后项不是顺着前项的意思说下去,而是突然转成同前项的意思相反或相对的说法,如例 2.4.8。

例2.4.8 a 中国尚无具体统计,|[转折]b{但}中国糖尿病人数正以每年七十五万新患者的速度递增。(chtb_0057)

让步关系,子句间具有让步转折关系,前一子句用表示让步的词语,表示先让一步,预示后面将有转折,如例 2.4.9。

例2.4.9 a{尽管}她的动作潇洒自如,|[让步]b{但}难度无法与罗莉相比,||[因果]c<因此>只获得 9.875 分,夺得银牌。(chtb_0310)

● 因果类

反映各种各样的因果聚合(参考汉语复句、汉语句群和修辞结构理论),因果类子句组成成分间地位不平等,各关系都有广义的原因或结果因素。

因果关系,一部分表示原因,另一部分表示由原因引起的结果。前后项可以是原因也可以是结果,如例 2.4.10。

例2.4.10 a 但{因为}中国经常收支顺差,||[递进]b 而且有一千多亿美元的外汇储备,|[因果]c<所以>将不会象其他亚洲国家那样陷入危机。(chtb_0122)

条件关系,句中一部分提出某种条件,另一部分推导出相应的结果,如例 2.4.11。

例2.4.11 a 评论指出,{只有}尊重朝鲜的社会主义制度,|[并列]b<并>取消对朝鲜的不当制裁,|[条件]c 朝鲜半岛的和平与稳定{才}能得到保障。(chtb_0653)

假设关系,一部分提出假设,另一部分表示假设实现后所产生的结果或不因假设实现而改变的结论。如例 2.4.12。

例2.4.12 a{若}不是货车上那位重伤者爬出来告诉救援人员,|[假设]b 伤亡将会更为惨重。(chtb_0687)

目的关系,子句之间有行为和目的关系,行为具有主观意图性,如例 2.4.13。

例2.4.13 a 进出口银行决定先在日本取得信用评级是为进入国际资本市场融资创造作准备,|[目的]b<以便>扩大资金来源,||[并列]c<并>支持中国机电产品和成套设备出口。(chtb_0010)

背景关系,前项交代事件的时间、地点、场景、人物等背景,后项叙述事件的内容。如例2.4.14。

例2.4.14 a 大连是中国常驻外商最集中的城市之一,[背景]b 目前已有外商投资企业五千二百二十六家,|||并列]c 合同金额一百五十八点一亿美元,|||并列]d 协议外资金额八十二点七亿美元,|||并列]e 实际利用外资三十一亿六千万美元。(chtb_0026)

推断关系,是推理性因果关系,句中一部分提出理由或根据,另一部分是从理由或根据推出的结论,如例2.4.15是由因推果,例2.4.16是由果推因。

例2.4.15 a 这种芯片在信息通信以及数字化家用电器中应用前景十分广泛。[[推断]b 预计到2003年,全世界超大规模集成电路芯片市场将达到1万亿日元(约合75亿美元)。(chtb_0757)

例2.4.16 a 领导很看重他,[[推断]b{可知}他确实有能力。(《现代汉语》邢福义 P263)

● 解说类

反映各种总分或解释关系,各子句之间地位是平等的,是对某一事物的解释、说明(参考汉语句群理论),主要有解说关系、总分关系、例证关系和评价关系。

解说关系,一般指后项对前项或前项中的某些词的解释、说明、补充。如例2.4.17。

例2.4.17 a 中国进出口银行最近在日本取得债券信用等级AA-,[[解说]b 这是日本金融市场当前对中国银行的最高债券评级。(chtb_0010)

总分关系,分说句是对总说句的某些实体词的分开解说。分说句至少包含两句以上或是其中的某一方面。如例2.4.18。

例2.4.18 a 经过三年的试点,可转换债券的发行方式和特征被国内所接受,||| b 今后借债主体将根据自身特点和市场需求在可转换债券、债券及银团贷款等方面进行选择,优化融资结构;||| c 有序开展发行无追索或有限追索权的项目融资债券,ABS等,||| d 从而化减金融机构对外借债的压力,||| e 推进项目管理国际化、科学化、透明化。[[总分]f{总之},我们会积极利用国际上有益的各种融资方式,整体上降低我们融资的成本,||| g 化减我们的金融风险。(分说句包括两个方面)(chtb_0132)

例证关系,前后子句是被证句的关系与例句的关系,例句是对被证说句的证释,如例2.4.19。

例2.4.19 a 重大科技成果迅速转化为现实生产力,是这个开发区的突出特点。[[例证]b 由浙江医科院院长、中国科学院院士毛江森主持在世界上率先研究成功,并具有国际先进水平的甲肝减毒活疫苗,去年经卫生部批准正式投入生产和使用,||| c 目前该区生产此疫苗的普康公司已形成年产五百万人份的生产规模,||| d 这对有效地控制甲肝流行具有重大意义。(chtb_0011)

评价关系，评价句是对被评价句所陈述的内容的评价，一般表明其作用或地位，如例 2.4.20。

例2.4.20 a 取名为“倍顺”的两家便民超市今天在此间开张营业。[[评价] b 它标志着美国大型跨国集团必纯士公司占有六成股份的厦门福兰普利超市有限公司正式启动。
(chtb_0124)

2.4.2.2 篇章关系的判定

在进行篇章关系判定过程中，会遇到一些问题，导致无法准确判定。问题主要有主观和客观两个方面。针对不同的问题需要采取不同的解决方法。

● 篇章关系判定的难点

篇章关系判定中的标注工作是个人对于关系判定的结果，个人主观意志的差异就会造成结果的不同，如思考角度不同、个人对篇章关系理解的偏差等。

主观因素之外，客观因素不容忽视。首先对汉语篇章关系类别的研究是一个不断完善的过程，由于现有的研究结果有限，篇章关系类别及定义还不够完整，在具体的关系分析中就难以形成准确的判断。从上文对篇章关系辨析中即可发现并列关系是一个相对松散类别，在实际标注过程中很难把握，这就需要在大量的实践基础上丰富对其的认识。

例2.4.21 a 这位官员说，国家同时也要引导外资投向新技术、新产品以及中西部地区的医药项目，[[目的] b {以此}带动民族医药企业提高产品的技术含量，|| [并列] c <并>增强国家市场竞争力。
(chtb_0121)

如例 2.4.21，通过分析，抽取关键词得到“提高技术含量”和“增强市场竞争力”，从其结构表明子句之间的结构相同，符合并列关系的结构，就其表层结构关系来说是并列关系，但深层分析其语义整体来看这一语段，能分析出“提高技术含量”的目的是“增强国际市场竞争力”。

另外，篇章中各子句之间的逻辑语义关系多数都为确定的一种，但也存在多种关系的情况，这就给篇章关系的判定带来一定的困难。

● 判定方法和依据

篇章关系判断总的要求就是以各种关系定义为依据。关系分类是汉语篇章结构分析的重要内容，在篇章结构分析时，要先将篇章关系进行整理分析，这样才能更好的进行分析工作。对各种关系定义要理解并掌握，把握每种关系的特征和实质。在形式上篇章关系有连接词则关系容易判定，但实际操作过程中占绝大多数的是没有连接词

的隐式篇章关系,因此,根据关系定义来判定是根本方法。

如果没有连接词,那么就可以采用添加连接词的方式来确定子句之间的关系。例2.4.22“扰素已经上市”的结果是“为中国一点二亿乙肝病毒携带者带来福音”,所以可以添加表示因果关系的连接词。

例2.4.22 a 扰素已经上市,|[因果]b<从而>为中国一点二亿乙肝病毒携带者带来福音。
(chtb_0040)

如果在句中找出关键性的词语,需要找出其内在的联系,如例2.4.23。

例2.4.23 a 国家医药管理局一位官员说,从今年起要逐步增加对创新药物研究的资助,|[并列]b 国有资本要以高技术领域和创新医药产品为投资重点。
(chtb_0121)

对例2.4.23分析过程中,有可能理解为因果关系,但是找到两个句子的中心词,前子句为资助,后子句为投资,资助和投资存在着差异。说明两子句之间是相互区别的,是叙述事物的两个方面。所以在语篇中前后子句之间的关系的判定可以通过判定句中中心词的关系来解决。

类比的方法,上文中已说明“不是……而是”是并列结构,运用类比的方法对例2.4.24进行判别,a子句和b子句容易误判为选择关系,但是观察之后会发现前后项之间并没有选择关系,连接词“不是……而是”和例句中的“不应该……而应该”类似,应该属于同一种关系。

例2.4.24 a 报告说,{不应该}把中国经济的高速增长看成是一种威胁,|[并列]b{而应该}视之为有助于促进亚太地区充满活力发展的积极因素。
(chtb_0166)

篇章关系判定过程中,标点对于关系判定也有很重要的作用,尤其是在并列关系中分号起到的标记作用。分号的出现,表示前后两子句间是一种并列关系,添加的连接词一般是表示并列的“并且”、“同时”等。这也是由分号的作用决定的,分号本身就表示前后两者的并列关系。

例2.4.25 a 统计资料显示,过去五年广西对外贸易和利用外资规模迅速扩大,|[解说]b 进出口贸易额累计达到一百亿美元,|||总分]c 其中出口六十八点七亿美元,||||并列]d 分别比“七五”时期增长一点七八倍和一点四三倍;|[并列]e 实际利用外资累计达到三十三点二四亿美元,|||并列]f 占改革开放以来累计总额三十八点三九亿美元的百分之八十四点四;|[并列]g 边贸成交额一百二十四亿元。
(chtb_0008)

篇章中的连接词具有标记关系的作用,除了连接词之外,有些词语也对关系判定有辅助作用。例2.4.26 A)中“预计”一词表明两子句之间是推断关系。B)中“这些”来指代上文中分说的各个因素,因此前后项是“分总”关系。

例2.4.26

A) a 据中国海关统计, 一九九五年两国贸易额已达一百六十九点八亿美元, |||并列|b 比前年增长百分之四十四点八。|[推断]c 经济专家预计, 今年中韩两国贸易额将增至二百五十亿美元。(chtb_0014)

B) a 证券分析员说, {由于}一九九六年香港地产大幅升值, |||因果|令深港两地地产价格差距拉大, |||并列|随着回归临近, 深圳地产开始回升, |||并列|而在深港联运方面也有较大发展空间, |[因果]这些因素为“深业控股”提供了良好的盈利环境。(chtb_0147)

2.5 篇章单位主次

连接词所连接的篇章单位, 根据重要性可区分为主要篇章单位和次要篇章单位。图 2.3 中, 箭头所指向的内容为主要篇章单位。篇章单位重要性的判断离不开篇章全局, 单纯一个关系中的两个篇章单位往往难于判断哪个重要。比如因果关系, 可能是“结果”重要, 也可能是“原因”重要, 究竟哪个重要需要根据全局重要性判断。篇章单位主次的一般判断标准是: 能代表所在整体与外界发生关系的篇章单位为主要篇章单位。

2.5.1 篇章单位主次区分

主次篇章单位的区分还与语序有关, 通常汉语的主要篇章单位在后, 次要篇章单位在前(见图 2.2、图 2.3 中主次篇章单位与语序的关系)。同一篇章关系中, 篇章单位的主次地位可能因语序变化而发生变化。比如因果关系中, 因果项语序颠倒后, 可能导致原因项成为主要篇章单位。

主次篇章单位的区分, 对于篇章结构分析应用于自动文摘等系统有重要作用。例如, 如果仅选一个句子作为例 2.1.3 的摘要, 最大可能是选择图 2.2 最低层、最右端的子句: “领导总是把一些重要的任务交给他”。

一个汉语篇章关系一般包含两个篇章单位, 这两个篇章单位是在同一个关系层中划分出来的。例如因项和果项是根据因果关系划分出来的篇章单位, 而因果关系层的中心是这两个篇章单位中能概括它所在关系层主旨的一个篇章单位, 又叫做单中心。但是在并列关系中, 篇章单位可以有多个, 并列关系的中心可能会由一个或多个篇章单位来充当, 即并列关系的中心可能是单中心, 也可能是多中心。

例2.5.1 四川鼓励外国公司参与股份制改革

a 据介绍,四川省计划近期内开始充分利用资本市场以及省内上市公司的“壳资源”
 ||[目的]*b 对弱势企业进行并购和重组, |||[并列]*c 扩大优势企业的规模, |[并列]*d 省政府
 热忱欢迎并鼓励外国公司参与。(chtb_0708 第2段)

a 四川是中国西部经济发展和对外开放较好的一个大省, *[并列]*b 也是中国最早开展
 农村改革和企业改革的地区。(chtb_0708 第3段)

例 2.5.1 中篇章结构的中心用“*”表示,“*”在“|”后,这表明“省政府热忱
 欢迎并鼓励外国公司参与”是第2段第1层关系的中心,并且是单中心。第2段第1
 层关系的前篇章单位和后篇章单位是并列关系,但是根据后篇章单位能概括出这篇文
 章的主旨“四川鼓励外国公司参与股份制改革”,并且与下文联系更紧密,这一关系
 层的中心应该是后篇章单位。

“四川是中国西部经济发展和对外开放较好的一个大省”和“也是中国最早开展
 农村改革和企业改革的地区”也是并列关系,语义上是从两个不同方面对四川省经济
 发展的评价,并且分别与文章主题中“外国公司”和“改革”相照应,因此这两个篇
 章单位都是这一关系层的中心,属于多中心。

汉语篇章中,一般情况下,每个关系层总是存在一个句子或子句能够概括它所在
 关系层的主要意思并且与上下文联系较紧密,能照应整篇的主题,这个句子或子句便
 是篇章结构的中心。从汉语篇章单位主次的定义来看,它有三大特点和功能:

第一,能够概括它所在关系层的主要内容或意图。如例 2.5.1 中“省政府热忱欢
 迎并鼓励外国公司参与”是第1关系层的结构中心,它把整段话的主要内容——“四
 川鼓励外国公司参与股份制改革”概括了出来。

第二,与上下文联系紧密。如例 2.5.1 中,第2段第1层关系的结构中心“省政
 府热忱欢迎并鼓励外国公司参与”与下文“四川是中国西部经济发展和对外开放较好
 的一个大省”的联系紧密程度比“四川省计划近期内开始充分利用资本市场以及省内
 上市公司的‘壳资源’,对弱势企业进行并购和重组,扩大优势企业的规模”与下文的
 紧密。

第三,符合文章的主旨。如例 2.5.1 中,第2段第1关系层的结构中心“省政府
 热忱欢迎并鼓励外国公司参与”与 chtb_0708 篇“四川鼓励外国公司参与股份制改革”
 的主旨相符合。

2.5.2 篇章单位主次判定

在判定篇章结构中心时,判定结构中心的原则有局部原则和全局原则。局部原则是指能够概括它所在关系层的主要内容或意图,并且与上下文联系紧密,这两个条件是缺一不可的,主要适用于比第1层关系低的关系层;全局原则指判定的篇章单位均符合局部原则时,需要结合文章的主旨来判定,此方法主要适用于第1关系层。

例2.5.2 新中国方志修纂体系初步建立

a 改革开放二十年间,中国的方志修纂工作出现了前所未有的繁荣景象。

(chtb_0710 第1段)

a 据中国历史文献研究会副会长、知名方志学家、浙江大学教授仓修良介绍,中国在不到二十年的时间里就出版了新修志书三千多种,*[[解说]b 其中新修的县志就达一千九百多种,*[[并列]*c 并在全国形成了拥有二万多专职修志人员、四万多名兼职修志人员的编纂队伍。

(chtb_0710 第2段)

a 中国的方志修纂已经有二千多年的时间。*[[并列]b 方志具有“存史、资治、教化”的功能,*[[因果]为历代统治者所重视。

(chtb_0710 第3段)

在例2.5.2中“中国在不到二十年的时间里就出版了新修志书三千多种”是第2段第2关系层的中心符合概括它所在关系层主要内容和紧密联系上下文的要求;而“其中新修的县志就达一千九百多种”从“其中”这一连接词得知“新修的县志”是包含在“新修志书”中的,不具有概括性。这一关系层中心的判定符合局部原则。

“中国的方志修纂已经有二千多年的时间。”和“方志具有‘存史、资治、教化’的功能,为历代统治者所重视。”均提到方志,都和文章主旨“新中国方志初步建立”有关,前篇章单位是从时间来介绍方志,后篇章单位是从功能来介绍方志。但是从上下文的联系紧密程度来看,方志的功能则显得不太重要。“中国的方志修纂已经有二千多年的时间。”侧面隐含了虽然修纂时间之长但是并未建立“修纂体系”,这一篇章单位和文章主旨更贴近。这一关系层中心的判定符合全局原则。

2.5.2.1 主次判定的依据

篇章单位主次的判定并不像现代汉语复句,依据句子间的关系来判定复句中心。在现代汉语中并列复句和顺承复句是没有中心句的,而在篇章结构这次判定中不存在此类情况。判定篇章结构中心并不是依据篇章关系来判定的,主要依据上下文和整篇文章的主旨。

例2.5.3 中国京杭大运河经济带迅速崛起

a 古老的京杭大运河如今{不仅}在贯通南北运输方面发挥重要作用, [[并列]*b{而且}带动起一条欣欣向荣的工业走廊, [[[因果]*c 形成了大运河经济带。(chtb_0013 第1段)

a 京杭运河古来繁华, [[[并列]b 两岸商贾云集, [[[并列]c 贸易发达。[[背景]*c 随着中国对大运河的整治, 运河航道状况得到极大改善, [[[因果]*d 许多企业纷纷看好这条“黄金水道”, *[[[解说]e 积极在此投资建厂, *[[[并列]f 沿河企业星罗棋布。*[[例证]g 据初步统计, 仅运量在万吨上的企业就有二千多家, [[[递进]*h 小的乡镇企业则不计其数。

(chtb_0013 第2段)

在例 2.5.3 中“带动起一条欣欣向荣的工业走廊”和“形成了大运河经济带”这两个子句的句间关系是并列关系, 是没有中心句可言的。但是在篇章结构中心判定中, 从这两个子句与这篇文章的主题“中国京杭大运河经济带迅速崛起”联系程度来说, “形成了大运河经济带”与主题联系紧密, 照应了主题; 从上下文的联系程度来说, “形成了大运河经济带”与下文“京杭运河古来繁华, 两岸商贾云集, 贸易发达。”联系比较紧密。更重要的是“形成了大运河经济带”这一子句能概括这一整段的意思。因此对篇章结构中心的判定并不是依据句间关系来判定的, 而是依据上下文和整篇文章的主旨。

篇章关系对篇章结构中心的判定也是有一定影响的。比如两个篇章单位之间的篇章关系是例证关系, 必然有一个是结论项, 一个是例子项。例子项是为结论项服务的, 是为了补充说明论点的。因此, 在例证关系的两个篇章单位中结论项是中心。

例2.5.4 a 北仑港的兴起, 还带动了当地一些地方企业的发展。*[[例证]b(如)宁波海天机械制造有限公司原来是一个不起眼的作坊, [[[对比]*c<而>他们借助港口的便利条件, *[[[并列]*d<并>不断开发出口新品进入美国、西班牙、希腊、澳大利亚等二十多个国家, [[[因果]*e<因此>目前已成为中国特大注塑机专业制造公司, *[[[并列]*f<并列>列中国机械部大中型企业综合指数百强之首。(chtb_0074 第4段)

例 2.5.4 中第一关系层中的两个篇章单位之间是例证关系, 由“如”字引出的例子项目的是为了说明“北仑港的兴起, 还带动了当地一些地方企业的发展”这一论点, 因此, 该论点便成为第一关系层的中心。

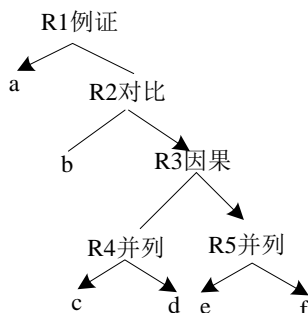


图2.13 例 2.5.4 篇章关系和中心树状图

不仅篇章关系会影响篇章结构中心的判定,而且相邻的两个关系层中上一层关系对下一层关系中心的判定也有一定的影响,即父关系对子关系中心的判定有一定的影响。将例 2.5.4 的篇章关系和中心转化成树状图更容易明白父关系对子关系中心判定的影响。例 2.5.4 的篇章关系和中心树状图如图 2.13 所示。

图 2.13 中结论项 a 和例子项 bcdef 属于第一层关系,它们之间的关系 R1 是例证关系,篇章单位 b 和篇章单位 cdef 属于第二层关系,它们之间的关系 R2 是对比关系。R1 和 R2 相比,R1 是父关系,R2 是子关系。在第一层关系中结论项 a 是中心,第二层关系的中心必须是和 a 的联系程度更紧密的篇章单位。转折前项 b 和转折后项 cdef 相比,b 更能说明论点 a。在这里,父关系的中心 a 决定子关系的中心。

判定一段话中每一层关系的中心是需要客观依据的,但写作者主观性的侧重和连接词也是判定汉语篇章单位主次的依据。

在判定篇章结构单位主次时,写作者的意图也是很重要的。

例2.5.5 a 少年姓孙, *|||并列]*b 属马, [*|并列]*c 比小水小着一岁, *|||并列]*d 个头也没小水高, |||转折]*e 人却本分实诚。(《浮躁》贾平凹)

从结构形式和逻辑意义角度出发,例 2.5.5 是在姓名、属相、年龄、外表、内在品质等方面对这个少年做一叙述,从这个层面上讲,这些子句都应是并列的,是同等重要的。但是作者用了“也”、“却”这两个连接词,可见作者主观上的侧重点是在第一层关系的后篇章单位 cde 和右边第二层关系的后篇章单位 e,因此“比小水小着一岁,个头也没小水高,人却本分实诚。”成为第一关系层的中心;“人却本分实诚”成为右边第二层关系的中心。

连接词对句子之间或子句之间的关系有一定的提示作用,它对汉语篇章层次主次的判定也起到相对重要的作用。

例2.5.6 过去的几十年里, (虽然)住房一直是中国农村居民的头等消费大事, ||转折]* (但)

在城镇,住房却被当作职工福利,*[[例证]由单位无偿分配,*[[并列]每月只需交纳极低的租金。

(chtb_0143)

例 2.5.6 中的复合连接词“虽然……但”在句中不仅起着连接两个子句的作用,而且揭示前后两个篇章单位是转折关系,并且连接词“但”和“却”有强调的作用。这表明“但在城镇,住房却被当作职工福利,由单位无偿分配,每月只需交纳极低的租金。”是这句话要表达的主要意思,并且也和这篇文章主旨中的“福利住房”有关联。因此可以判定第一关系层的后篇章单位是中心。

2.5.2.2 主次判定的方法

判定汉语篇章层次中心的方法大致有比较法和删除法两种。

比较法是将判定的篇章单位放到整段或是整篇文章中,比较各个篇章单位与上下文和文章主旨的联系紧密程度,联系紧密程度大的篇章单位就是它所在关系层的中心。

例2.5.7 a 伊拉克国防部官员说,多国部队的飞机 13 日当地时间凌晨 4 点投下的两枚导弹或炸弹,击中一座大型地下防空掩体,[[因果]*<因此>躲在里面的至少 500 名平民被炸死,*[[总分](其中)多半是妇女和儿童。

(chtb_0179 第 4 段)

例 2.5.7 用的是比较法,将第一层的前篇章单位“多国部队的飞机 13 日当地时间凌晨 4 点投下的两枚导弹或炸弹,击中一座大型地下防空掩体,”和后篇章单位“躲在里面的至少 500 名平民被炸死,其中多半是妇女和儿童。”放到整篇文章中,用这一方法才能看出后篇章单位和文章主旨联系紧密,并且和上下文联系紧密。

例2.5.8 a 突尼斯再次呼吁国际良知采取行动,停止流血,结束毁灭性战争,*[[并列]b(并)为和平解决冲突创造条件。

(chtb_0179 第 5 段)

例 2.5.8 中,“并”这一连接词提示了第一层前后两个篇章单位是并列关系,但在语义上却是有轻重之分的,这时可以用删除法。如果删除第一层关系的后篇章单位 b 时,上下文衔接得仍然紧凑,同时又不损害整篇文章的主旨大意。如果删除前篇章单位 a 的话,语义上跳度太大,上下文衔接也不顺畅。因此后篇章单位 b 才是这句话的结构中心。

在短时间内无法判断哪一个篇章单位是它所在关系层的中心时,删除法是一个快捷的方法,不过此方法最好和比较法结合使用。

例2.5.9 富士山游吟

a 五年一次的国际日耳曼语言文学学会年会今年适逢在东京举行, *[[因果]b 这{使}笔者得以一睹富士风采, *[[解说]c 可算实现了一桩多年的“非分之想”。

(chtb_0207 第2段)

a 富士山离东京只有 80 公里——*[[解说]b 这显然又是造化的特意安排: *[[解说]c 这个岛国最高的山与她的最大的城市相依为伴, *[并列]d<并>互为辉映。*[顺承]*e(于是), 轻舟熟路, f[[因果]*f 我们乘坐的旅游大轿车径向西南——富士山方向疾驶。

(chtb_0207 第3段)

a 不知不觉间浓荫几乎完全封死了马路的上空, *[[因果]b 原来马路已经变窄了, *[[并列]*c(并且)有了坡度。

(chtb_0207 第4段)

例 2.5.9 中第 3 段第一关系层的前篇章单位和后篇章单位是顺承关系, 这从“于是”这一连接词可以看出。如果利用删除法, 将第 3 段第一关系层的前篇章单位删除时, 损害了顺承连接词“于是”所连接的前后两个篇章单位时间上先后相承的关系。反之, 删除后篇章单位时, 语义仍然完整。这样就可能会判定前篇章单位是这一关系层的中心。但是删除后篇章单位“于是, 轻舟熟路, 我们乘坐的旅游大轿车径向西南——富士山方向疾驶。”过渡到下一段却显得太突兀, 损害整篇文章的完整性。仅用删除法来判定结构中心, 会对文章的语义完整性带来损失, 删除法需要和比较法结合使用。

2.5.2.3 主次判定的难点

篇章单位主次判断比较困难的情况主要有: 同一关系层的各个篇章单位之间的关系不止一种时, 即存在多个关系; 同一关系层的各个篇章单位之间存在误导的连接词时, 即连接词显示的关系与篇章单位之间深层的语义关系不一致; 同一关系层各个篇章单位的内容在文章的上下文中均涉及到, 并且篇章单位顺序不符合常规时, 判定结构中心会受到阻碍。

在例 2.5.7 中第二关系层的前后两个篇章单位存在多个关系。这两个篇章单位是由连接词“其中”连接着的, 但这一关系层的中心很难判定。“其中”一般表示其连接的前后两个篇章单位是整体与部分的关系, 但文中的连接词“其中”也暗含前后两个篇章单位是等价的。因此即使有连接词“其中”也很难判定哪个子句是这一关系层的中心。可以结合比较法和删除法来解决这一问题。先删除例 2.5.7 第二层关系前篇章单位和后篇章单位中的任何一个, 比较它们和这段文章的联系紧密程度。前篇章单位“躲在里面的至少 500 名平民被炸死”在语义上比后篇章单位“其中多半是妇女和

儿童”更适合做第一层关系的结果项。最重要的是后篇章单位和文章的主题“德奎利亚尔对伊平民死亡表示悲痛一些国家谴责多国部队屠杀无辜”联系更紧密。因此解决这一问题可以根据同一关系层各篇章单位与上下文和文章主题的联系程度来判定这一关系层的中心。

当遇到不能准确标示层次关系的连接词,在判定汉语篇章的层级中心时便会产生分歧。例 2.5.8 中的“并”在这一句话中是连接词,表面上是表示并列关系的,但是深层语义上“并”前后的两个篇章单位是有轻重之分的,因此产生了分歧。这可用删除法同时结合整篇的文章主旨大意的方法来判断这句话中的哪一个篇章单位更能表达这一关系层的主要意思并且与上下文连接紧密。

当同一关系层各个篇章单位的内容在文章的上下文中均有涉及到,并且篇章单位不符合常规分布时,这一层关系的中心也很难判定。如例 2.5.10:

例2.5.10 a 昨天晚上,我国驻法大使馆一派节日气氛。[[解说]*b{为}迎接新春佳节,[[目的]*c

蔡方柏大使举行盛大酒会,*[[[目的]d<以>款待 500 多位侨居法国的侨胞。

(chtb_0183)

例 2.5.10 中第三关系层的中心,因两个子句的内容在下文中都有涉及,很难判定。这时便要着眼于整篇文章,看哪一个篇章单位更能代表它所在关系层的主要意思,并与上下文衔接紧密。

2.6 与相关理论的比较

从 1.2 节的介绍可知,目前的篇章结构理论研究最具代表性的是 RST 和 PDTB 体系,下面将本文基于连接依存树的汉语篇章结构表示体系和它们进行简单的对比。对比结果如表 2.1 所示。

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。

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。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。

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。

表2.1 CDT 与 RST 和 PDTB 体系对比

类别	RST	PDTB 体系	本文 CDT 表示体系
基 本 篇 章 单 位	EDU, 定义明确, 一个关系有一个 或多个 EDU	论元; 谓词-论元模 式; 一个关系有两个 论元	子句; 从 3 方面进行明确定义; 是 自顶向下切分的终点; 一个关系有 两个或多个子句
连 接 词	--	标注显式连接词和添 加隐式连接词	显式连接词和隐式连接词; 显式连 接词是否可删; 添加的隐式连接词
篇 章 关 系	给定语义类别并 标注	给定三层语义类别; 标注语义类别和连接 词	用连接词代表关系; 标注连接词及 其属性; 将连接词映射到一个三层 的关系体系上
结 构	完整篇章结构树	从连接词及论元标注 中可推导部分树	完整篇章结构树; 自顶向下分割; 结构可用连接词的关系层次表示
主 次	主次由具体关系 类别决定	--	主次由全局意图决定, 跟关系无直 接关系

在篇章单位主次区分上, PDTB 体系不区分主次, RST 严格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 借鉴了 RST 和 PDTB 的优点并结合了汉语本身的特点。

2.7 本章小结

本章提出了基于连接依存树的汉语篇章结构表示体系, 该体系借鉴汉语复句、汉语句群、英语 RST 和 PDTB 体系的研究成果。基于连接依存树的汉语篇章结构表示体系主要包括子句、连接词和篇章结构关系、篇章单位主次几个关键元素。该体系从汉语特点出发定义子句, 子句位于连接依存树的叶子节点。定义起连接作用的词为连接词, 连接词位于连接依存树的中间节点, 连接词本身既可以表示逻辑语义关系, 其

在连接依存树中的层次又可以表示篇章结构。针对隐式篇章结构较难处理的现状,采取区分显式连接词是否可删以及添加隐式连接词的方法进行处理,并给出切实可行的连接词添加和删除方法。篇章单位主次不是采用简单根据篇章关系判断,而是采用全局性标准,即能代表所在关系整体与外界发生关系的为主要篇章单位。通过大量实例及针对实例的具体操作说明本体系是切实可行的,且充分体现了汉语篇章结构的特点。

本章基于连接依存树的汉语篇章结构表示体系部分内容发表在自然语言顶级会议 EMNLP2014 上,子句定义与判断内容整理发表在 EI 检索的词汇语义学会议 CLSW2012 上。

第3章 基于连接依存树表示体系的 CDTB 语料库构建

本章主要是以连接依存树表示体系为理论基础,选取一定语料,按照标注规范和流程,通过标注平台,创建了汉语篇章结构语料库(Chinese Discourse Treebank, CDTB)。首先介绍 CDTB 语料库标注策略、标注方法等,然后给出语料标注一致性测试结果,最后对 CDTB 语料的主要标注内容进行统计分析。

3.1 引言

近年来,统计自然语言处理异军突起,现已成为自然语言处理的主流。统计自然语言处理的主要需求包括计算机、语料库和软件^[88]。语料库就是一个由大量在真实情况下使用的语言信息经过科学的收集和组织而集成的专供研究使用的资源库。语料库并非语篇的简单堆砌或集合,它的主要特征是样本具有代表性、规模有限、语料文本以电脑可读的形式存在(以文本文件或 XML 文件格式存储)^[89]。语料库对于自然语言处理研究的巨大价值已经得到越来越多的学者认可。

语料库标注就是为语料库增加一些语言学信息,有时称为语料库加工。最常见的一种标注就是在原始文本中增加空格将汉字分词,这就是所谓的分词标注^[90]。除分词外,根据文本分析的不同层次,还有词性标注(根据事先定义好的词性类别为分词添加词性,如 Brown 语料库、北京大学《人民日报》语料库)、句法标注(按照不同的句法分析理论或句法分析模型给文本中的每个句子增加句法信息,如宾州树库)、语义标注(在原始语料中增加语义信息,语义可以是词义、句义和篇章义等信息,如 WordNet)、篇章标注(就是为原始语料增加篇章信息,如 RSTDT、PDTB)等。

语料库标注的常用方法有传统的以人工为主的标注方法和半自动标注的方法。半自动标注一般是先用自动标注工具进行初标,然后用人工校对的方法建立语料库。但对于篇章结构语料标注来说,由于需要标注的内容比较复杂,这种半自动的标注方法在实际标注中很难被采用,所以采用人工方法建设篇章结构语料库是目前比较通行的方法。

3.2 自顶向下的 CDTB 标注策略

本章主要进行汉语篇章结构语料库构建,在第2章所提出的基于连接依存树的汉

语篇章结构表示体系基础上,我们进行了标注人员培训并制定了相应的标注策略。

CDTB 语料的标注工作由作者和一位语言学博士指导核对,四个汉语言文学的本科生进行具体标注。四个学生分成 2 组进行标注,整个标注过程分为 4 个阶段。第 1 阶段(3 个月),由于语料库加工工作量很大,为保证质量,也为保证通用性,根据本文提出的基于连接依存树的汉语篇章结构表示体系,作者制定了初步的标注规范,开发了汉语篇章结构标注工具,并对 4 个学生进行了连接依存树表示体系和所开发标注工具使用的培训。第 2 阶段(4 个月),为保证标注一致性,所有标注者分别标注相同的 50 个文档(大约 260 个段落),然后在一起逐一校对讨论,讨论内容涉及子句、连接词、篇章关系、篇章单位主次、篇章结构层次等所有的标注内容,最后统一标注思想,得到修订的标注规范。第 3 阶段(9 个月),4 个标注人员分成 2 组完成 450 篇文档的标注,其中有 60 篇 2 组标注人员都分别进行了标注,这个数据主要是用来计算 CDTB 语料标注的一致性。标注规范的制定和语料库的标注实践,两者形成一个循环,必须经过多个回合的研究与实践才能逐步完善,第 3 阶段完成后,形成最终的标注规范。第 4 阶段(3 个月),根据最终的标注规范,作者对所有的标注语料都逐一进行校对,校对内容包括子句、连接词、篇章关系、篇章单位主次和篇章结构层次等。

CDTB 标注采用了自顶向下的标注策略,即对每一段内容先找出其最上层关系,然后递归的对切分后的内容进行标注。使用自顶向下的主要考虑是:第一,这种策略有利于宏观上把握整体结构,符合篇章分析所具有的全局整体性特征。第二,由于汉语句子和短语结构没有明显形式区别,这种策略一定程度上避免了汉语子句切分困难对篇章结构分析的干扰。在这种自顶向下分析策略中,子句切分成为一个末端问题,即使子句切分出现偏差,影响也是局部的,不至于对篇章整体结构造成大的影响。对于汉语篇章结构分析,自顶向下的分析策略实则先易后难,它不仅可以避免错误,控制影响范围,而且可以提高标注效率。第三,这种策略符合汉语习惯,在汉语句群和汉语复句分析中常采用,也比较符合汉语篇章理解的一般心理过程。当然,我们并不排斥自底向上方法,在某些篇章局部,自底向上和自顶向下可以相互结合使用。

3.3 人机结合的 CDTB 标注方法

CDTB 标注时首先任意给定一个篇章作为需要标注的生语料,然后利用汉语篇章结构标注平台对语料进行切分子句、指定连接词、标注连接词属性、标注关系类型、

指定中心句等操作生成一个个关系,根据各个关系的标注信息将一个篇章中的所有关系组合在一起构成一棵连接依存树。为保证标注一致性,还需要对多个标注者使用此标注平台产生的标注语料进行一致性计算。标注完成后还需要对标注结果进行统计分析。直接在生文本上进行手工标注非常不直观且易出现错误,为更有效的进行实际标注,本文开发了汉语篇章结构标注平台,下面介绍标注平台使用流程及语料保存结果。

3.3.1 标注流程设计

结合所需功能,标注平台的工作流程如图 3.1 所示。首先分析语料,其次标注某个关系的子句、连接词、连接词属性、连接词类型和篇章单位主次等信息,根据标注的关系层次画出关系树,直至所有关系标注完毕生成此文档的整体篇章结构树,然后将标注结果存储为 XML 文件,此 XML 文件即为输出语料。对生成的 XML 文件进行统计分析,得出统计结果。对多个标注者的标注结果进行一致性计算,得到多个标注者标注的一致率。

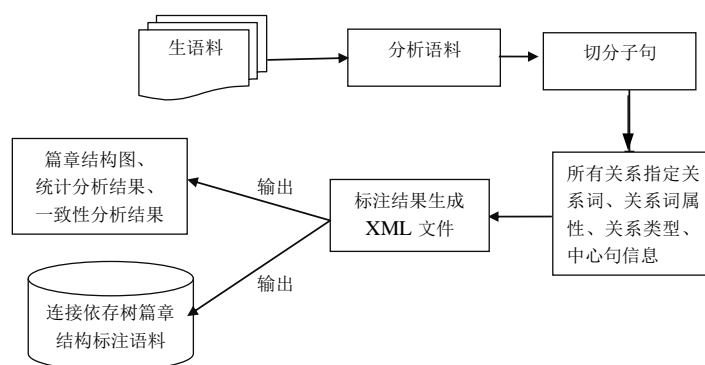


图3.1 汉语篇章结构标注平台处理流程

根据所需功能及处理流程,本文开发了基于连接依存树的汉语篇章结构语料标注平台。主要功能有语料标注、语料统计和语料一致性计算。其中语料标注是主要内容,包括新建关系、保存关系、修改关系、删除关系等操作,每个关系有关系类型、连接词、子句等属性可选择或添加。

3.3.2 语料标注

标注语料开始前需要选择未标注文件所在的主目录,需要标注的文件名,标注后文件的保存路径,是否显示标注过程中的篇章结构关系图。下面以例 2.1.4 为例介绍标注方法。

首先载入内容“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，因此大量出现的是以前不曾遇到过的新情况、新问题。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，而是借鉴发达国家和深圳等特区的经验教训，聘请国内外有关专家学者，积极、及时地制定和推出法规性文件，使这些经济活动一出现就被纳入法制轨道。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，运转至今，成交药品一亿多元，没有发现一例回扣。”。

标注语料主要工作是标注篇章关系及所对应的属性。篇章关系可以进行新建、添加、修改和删除操作，基本功能大致相同，下面以新建为例进行说明。新建篇章关系采用自顶向下的方法进行，即首先标注最上层的关系。

层次用特殊的符号（如“|”）进行标注。一个“|”代表是第一层的关系， n 个“|”就代表第 n 层关系。如例2.1.4可以这样标注：

例2.1.4中第1层关系标注为：连接词是“因此”；子句1是“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，因此大量出现的是以前不曾遇到过的新情况、新问题。”，子句2是“对此，浦东不是……没有发现一例回扣。”；连接词属性是不可删除的连接词；关系类型是显式关系；属于背景关系；中心句是子句1。

对第1层标注完成的左右部分分别进行第2层标注，例2.1.4中第2层有2个关系：第1个关系：连接词是“因此”；子句1是“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，”；子句2是“因此大量出现的是以前不曾遇到过的新情况、新问题。”；连接词属性是可删除的连接词；连接词类型是显式关系；中心句是子句2。第2个关系：连接词是“例如”（添加的连接词）；子句1是“对此，……被纳入法制轨道。”；子句2是“去年初浦东新区……没有发现一例回扣。”；连接词属性是可添加的连接词；连接词类型是隐式关系；关系类别是例证关系；中心句是子句1。

对第2层标注结构进行第3层次的标注，例2.1.4第3层有2个关系，第1个关系：连接词是“不是……而是”；子句1是“对此，浦东不是……的做法，”；子句2是“而是借鉴发达国家……被纳入法制轨道。”；连接词属性是不可删除的连接词；连接词类型是显式关系；关系类别是并列关系；中心句是子句1和子句2。第2个关系：连接词是“正因为”；子句1是“去年初浦东……正因为一开始就比较规范，”；子句

2 是“运转至今……没有发现一例回扣。”；连接词属性是不可删除的连接词；连接词类型是显式关系；关系类别是因果关系；中心句是子句 2。

例 2.1.4 第 4、5 层次参考前面的方法从顶向下依次进行标注，直到篇章全部切分为子句停止。

3.3.3 语料格式

语料标注结果保存在 XML 文件中，一个关系对应一行记录，例 2.1.4 标注的部分内容如图 3.2 所示：

```
<P ID="2">
  <R ID="1" StructureType="False" RelationType="显式关系" Layer="1" Connective="对此"
  ConnectiveType="解说" ConnectivePosition="61...62" ConnectiveAttribute="可删除" Centre="2"
  LanguageSense="True" General="True" Sentence="浦东开发.....新问题。|对此.....回扣。"
  SentencePosition="1...60|61...230" ChildList="2|3" ParentId="-1" UseTime="87" />
  <R ID="2" StructureType="False" RelationType="显式关系" Layer="2" Connective="因此"
  ConnectiveType="因果" ConnectivePosition="37...38" ConnectiveAttribute="可删除" Centre="2"
  LanguageSense="True" General="True" Sentence="浦东开发.....工程，|因此.....新问题。"
  SentencePosition="1...36|37...60" ChildList="" ParentId="1" UseTime="83" />
  <R ID="3" StructureType="逐层切分" ConnectiveType="隐式关系" Layer="2"
  RelationNumber="单个关系" Connective="例如" RelocationType="例证关系"
  ConnectivePosition="169" ConnectiveAttribute="可添加" Rolelocation="normal"
  LanguageSense="true" Sentence="对此，.....被纳入法制轨道。|去年初浦东新区.....没有发现
  一例回扣。" SentencePosition="61...167|168...230" Center="1" ChildList="4|5" ParentId="1"
  UseTime="120"/>
  <R ID="4" StructureType="逐层切分" ConnectiveType="显式关系" Layer="3"
  RelationNumber="单个关系" Connective="不是...而是" RelationType="并列关系"
  ConnectivePosition="66...67|100...101" ConnectiveAttribute="不可删除" RoleLocation="normal"
  LanguageSense="true" Sentence="对此，.....的做法，|而是借鉴发达国家.....纳入法制轨道。"
  " SentencePosition="61...99|100...167" Center="2" ChildList="6" ParentId="3" UseTime="41"/>
  .....
  <R ID="6" StructureType="True" RelationType="隐式关系" Layer="5" Connective="并"
  ConnectiveType="并列" ConnectivePosition="" ConnectiveAttribute="可添加" Centre="3"
  LanguageSense="False" General="True" Sentence="而是.....教训，|聘请.....学者，|积极.....文件，"
  " SentencePosition="100...119|120...131|132...148" ChildList="" ParentId="4" UseTime="26" />
  .....
</P>
```

图3.2 标注语料 XML 形式保存结果

XML 中每一段一块记录，用<P></P>界定。<P ID="2">表示第一段内容。每一段

中有一个或多个关系，用<R ……/>表示，其中：ID 表示这个关系的 ID 号；StructureType 表示节点类型，并列切分包含多个子句；ConnectiveType 是关系类型，有显式和隐式关系两种；Connective 是连接词，关系为显式关系时表示句子中存在的连接词，隐式关系表示手工添加的连接词；RelationType 是具体的关系类型，如“并列、转折、递进、因果”等；ConnectivePosition 记录连接词在段落中的位置；ConnectiveAttribute 是记录连接词是否可删除、可添加；Center 的值为 1 时表示它的中心句为左子句（子句 1），为 2 时表示它的中心句为右子句（子句 2），为 3 时表示它的中心句为它的所有子句；Sentence 里放的是它的所有子句的内容，以“|”分隔；SentencePosition 里放的是所有子句在段落中的位置信息，以“|”分隔，“…”前是子句开始位置，“…”后是子句结束位置；ChildList 里放的是它的所有孩子节点的 ID 号，以“|”分隔；ParentId 的值是它的父亲节点的 ID 号；UseTime 是建立这个节点时所花的时间（秒数）。通过以上的关系标注信息，即可构建一棵篇章结构树。

3.3.4 语料校对

由于本文所标语料要求每个篇章（段落）生成一棵树，因此需要对语料进行检查以保证标注结果可以生成一棵树，检查分为自动检查和人工检查。对标注语料利用树遍历程序看其是否可以构建一颗树，如不能构建则进行人工检查。人工检查主要是检查一棵树是否只有一个根节点，对树中的叶子节点进行检查看其是否覆盖整篇文档，如果没有遗漏则检查标注的属性信息如子句、连接词、篇章关系、篇章单位主次等标注是否正确。

对标注好的语料，为保证标注子句、连接词、篇章关系、篇章单位主次等判断的一致性，我们对标注好的所有语料分别进行了人工校对。校对时每个人只负责一个标注属性，如本人负责子句校对，学生甲负责篇章关系校对，学生乙负责连接词校对等。这样校对完成后可以尽量去除不同标注者的个性差异。

3.4 CDTB 标注一致性测试

标注一致性是衡量标注质量的重要标准。我们选取两名标注者 A 和 B 各自独立标注的 60 篇文档（chtb0001_0041—chtb0001_0100），进行一致性测试。一致性评估主要计算标注一致率，一致率(Agreement)= $A \cap B / A \cup B$ ，主要考察两名标注者标注的一致内容。对于不同的标注任务，其计算内容根据具体情况有所不同。严格来说，比

较两个标注者的输出来确定是否一致的方法还不足够,因为忽略了标注偶然一致的情况^[90],本文还对部分内容进行了 Kappa 值^[91]的计算,计算方法举例如下。

假设标注者 A 和标注者 B 共同标注子句是否可切分的信息。现在共有 100 个标点需要标注,其中 A 标注了 55 个标点为正例(可切分),45 个标点为负例。B 标注了 50 个标点为正例,50 个标点为负例。他们对同一个标点都标注为正例的个数为 45 个,都标注为负例的为 40 个。可以构建表 2.2 所示二维表。

$K(A) = X1+X2 = 0.85$, $K(A)$ 是标注一致的比例

$K(E) = Z1*Z2 + Z3*Z4 = 0.5$, $K(E)$ 是标注者偶然一致的比例

Kappa 值 = $(K(A)-K(E))/(1-K(E)) = 0.7$, Kappa 值是 AB 各自标注正例的概率相乘加上标注为负例的概率相乘

表2.2 Kappa 值计算二维表

标注者	类别	B		求和
		+	-	
A	+	0.45 (X1)	0.10	0.55 (Z1)
	-	0.05	0.40 (X2)	0.45 (Z3)
	求和	0.50 (Z2)	0.50 (Z4)	

如果需要计算多个类别,可以将表 2.2 扩展,仍然算对角线的概率为 $K(A)$, $K(E)$ 为 AB 分别标注为不同类别的概率相乘后加起来,如标注总共有三个类别,则 $K(A)$ 为 AB 标注类别完全一样的概率,即都标为类别 (1,1) (2,2) (3,3) 之和除以总数。 $K(E)=A$ 标类别 1 的概率*B 标类别 1 的概率+A 标类别 2 的概率*B 标类别 2 的概率+A 标类别 3 的概率*B 标类别 3 的概率。

根据以上一致率和 Kappa 值,计算 CDTB 标注一致性结果见表 2.3。在表 2.3 中,本文计算子句切分一致性的方法是判断两个标注者 A 和 B 相同切分的个数(交集)和他们所有切分的个数(并集)。对句子位置 $SentencePosition="X_1 \cdots X_2|Y_1 \cdots Y_2"$,计算 A、B 标注切分位置相同的情况。汉语子句的切分位置均有标点标记,将所有可能作为切分标记的标点(句号、分号、问号、逗号等)作为总数。子句切分一致率为 91.7%, Kappa 值为 0.84,这个结果表明本文子句定义是合理的,也是便于实现的。

显式隐式关系判断是对相同切分位置的标点,计算其都标注为显式或隐式关系的个数,表 2.3 显式隐式关系判断一致率为 94.7%, Kappa 值为 0.81,说明对于显式和隐式关系的判断正确率较高,因为识别显式关系和隐式关系有连接词作为标志。

表2.3 CDTB 一致性测试结果

类别	Agreement	Kappa
子句切分	91.7	0.84
显式隐式关系判断	94.7	0.81
显式连接词识别	82.3	-
隐式连接词插入	74.6	-
单中心或多中心判断	80.8	-
单中心主次判断	82.4	-
结构（相同连接词、相同子句）	77.4	-

对于显式关系，两个标注者选择相同显式连接词的一致率为 82.3%，这是因为本文在计算一致率时严格要求两个标注者选择的连接词完全一样，如一个标注者标注连接词为“也……并”，另外一个仅选择了“并”，则两者标注在严格标准下被认为标注不一致，如果我们将标准放松至两个标注者标注的连接词可以互相包含，则一致率可以达到 98%。

隐式连接词插入一致率指的是相同隐式关系，两个标注者在相同位置插入相同连接词的一致情况。表 2.3 隐式连接词插入一致率较低，为 74.6%。主要原因是对于同一隐式篇章关系，存在多种可以插入的连接词，如在表示因果关系的隐式关系中，可以插入“因此”、“所以”等词表示同一种关系。如果我们对同一种篇章关系可以插入的连接词进行限制，一致率可以达到 84.5%。

单中心或多中心判断是指同一关系，两个标注者都标为单中心或多中心的一致率，为 80.8%。单中心主次指对于同一单中心关系，两个标注者指定中心在前或在后相同的一致率，为 82.4%。由此可知，汉语中心判断一致率还有待提高，主要原因是本文根据篇章意图确定主次地位，而相同的内容，不同的人理解会有所不同。

最后，总的篇章结构一致率为 77.4%，所谓篇章结构一致指在相同位置，子句切分一致，子句范围一致，显式和隐式连接词也一致。这个结果说明对于本身歧义较大的汉语篇章结构，本文所提基于连接依存树的汉语篇章结构表示体系是合理的。

3.5 CDTB 标注信息统计与分析

目前 CDTB 共有 500 个文档（chtb001-chtb0657），全部来自 CTB 语料，在 CTB

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，每个子句平均长度为 22 个汉字。

下面分别从连接词、篇章关系、篇章结构、篇章单位主次等方面对 CDTB 进行详细的统计分析。

3.5.1 连接词统计与分析

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。单义连接词（可表示一种关系类别）有 246 个，多义连接词（可表示多种关系类别）有 35 个。不同的连接词可以表示不同的关系，具体连接词和关系对应表见附录 1。表 2.4 给出 CDTB 中出现次数最多的 10 个显式连接词和隐式连接词。

表2.4 CDTB 中出现次数前 10 的显式和隐式连接词

显式连接词				隐式连接词			
连接词	出现次数	连接词	出现次数	连接词	出现次数	连接词	出现次数
并	208	其中	154	因此	371	并	355
也	133	而	70	并且	257	例如	139
但	69	还	68	来	68	以	62
使	56	以	52	然后	55	其中	47
为	47	同时	46	而	47	因为	32

从表 2.4 可以发现，显式连接词和隐式连接词分布有一定的差别，显式连接词多是我们平常经常使用的词，如“也”、“但”、“还”、“为”等。隐式连接词是表示具体关系的词，如“因此”、“并且”、“例如”、“然后”、“因为”等。表 2.4 中有几个连接词分别作为显式连接词和隐式连接词出现，如“并”、“而”、“以”，这个结果也说明本文提出的显式连接词可删和隐式连接词可添是汉语的实际情况。

表2.5 显式连接词词性分布

词类	总数	连词	介词	副词	其它
个数	274	104	56	73	41
所占比例	100%	38.0%	20.4%	26.6%	15.0%

2.3 节提到, 连接词有不同的词性, 主要有连词、介词和副词, 每种词性的连接词分布如表 2.5 所示。由表 2.5 可知, 连词在所有连接词中所占的比例最大, 为 38.0%, 其次为副词, 占 26.6%。

由 2.3.1 可知, 连接词在形式上有独用 (如“虽然”)、关联词 (如“虽然……但是”)、合用 (如“同时也”) 和其它形式 (如“虽然……但……却……”)。CDTB 中, 独用连接词有 136 个, 占 49.6%, 关联词有 90 个, 占 32.8%, 这两类是连接词的主要形式。连接词合用的形式比较少, 只有 5 个, 其它形式的连接词有 43 个, 占 15.7%。

汉语篇章中的连接词, 有些通过连接词对句意关系的判断不明确, 会有同时认为是甲关系和乙关系的情况。一个连接词同时表示多种关系, 这类连接词称为多义连接词。虽然可以表示多种关系, 如连接词“并”可以表示“并列关系 (199)”和“顺承关系 (8)”, 但在标注的过程中, 一般可以根据连接词的上下文环境判断出是哪一种关系。这类连接词共有 35 个, 部分词及其所能表示的关系如表 2.6 所示。从表 2.6 可知, 连接词“而”可以表示“转折关系、递进关系、对比关系、并列关系、顺承关系、因果关系”共 6 种关系。多义连接词是篇章关系判断的难点, 虽然人工可以根据上下文环境进行区分, 但自动判断多义连接词的篇章关系仍没有有效的方法。

表2.6 部分同时表示多种关系的连接词

序号	连接词	同时表示的关系
1	而	转折关系、递进关系、对比关系、并列关系、 顺承关系、因果关系
2	否则	假设关系、转折关系、推断关系、条件关系
3	不仅……而且	并列关系、递进关系
4	如	例证关系、假设关系
5	或	选择关系、并列关系

3.5.2 篇章关系统计与分析

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%)。英语中显式关系和隐式关系基本相当^[65], 而汉语显式关系和隐式关系比重为 1: 3, 这表明汉语中隐式关系所占比例明显高于英语。除标注连接词外, 本语料对每个关系均标注关系类型, 由 2.4 可知, 本语料中的篇章关系类型也可以根据连接词对应关系推导出来, 也可以推导出其它分类。根据 2.4 的连接词分类方法, 各种关系类型统计见表 2.7 所示。

表2.7 各种关系类型个数统计表

大类	关系类型	显式	隐式	全部	百分比(%)
因果类	因果关系	204	482	686	9.38
	推断关系	3	35	38	0.52
	假设关系	55	14	69	0.94
	目的关系	163	170	333	4.56
	条件关系	37	34	71	0.97
	背景关系	4	130	134	1.83
因果类小计		466	865	1331	18.21
并列类	并列关系	742	2765	3507	47.98
	顺承关系	133	382	515	7.05
	递进关系	52	7	59	0.81
	选择关系	10	0	10	0.14
	对比关系	38	22	60	0.82
并列类小计		975	3176	4151	56.79
转折类	转折关系	157	39	196	2.68
	让步关系	16	0	16	0.22
转折类小计		173	39	212	2.90
解说类	总分关系	158	77	235	3.21
	例证关系	29	224	253	3.46
	评价关系	3	218	221	3.02
	解说关系	10	897	907	12.41
解说类小计		200	1416	1616	22.11
合计		1814	5496	7310	100

由表 2.7 可知,所有小类关系中并列关系个数最多,有 3507 个,占有所有关系的 47.98%。选择关系仅 10 个,这主要与本语料的来源有关,新闻语料通常是陈述事实,所以选择关系较少。四大类中,并列类有 4151 个,所占比例高达 56.79%,其中,隐式并列类有 3176 个,占有所有隐式关系的 57.79%。转折类仅有 212 个实例,占 2.90%,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。

3.5.3 篇章结构统计与分析

CDTB 中记录了每种关系所在的层次,层次为 1 的位于篇章树的最上层,层次为 2 的位于第 2 层,数字越大越靠近篇章结构树的叶子节点,具体关系层次分布见表 2.8。

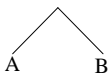
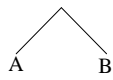

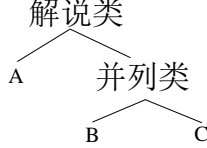
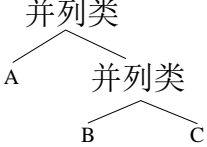
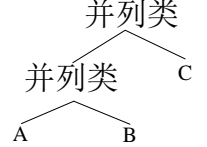
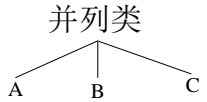
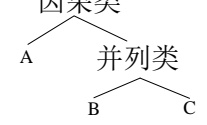
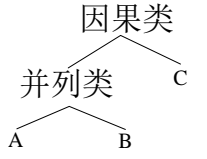
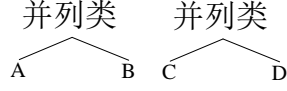
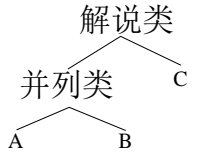
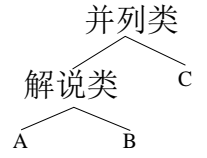
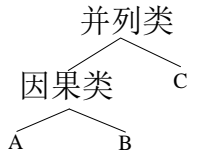
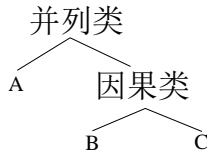
表2.8 CDTB 关系层次分布

关系类型	层次 1 (个数 比例)		层次 2 (个数 比例)		层次 3 (个数 比例)		层次 4 (个数 比例)		其它层 (个数 比例)		合计
背景关系	98	73.13	26	19.40	5	3.73	5	3.73	0	0.00	134
并列关系	887	25.29	1184	33.76	842	24.01	382	10.89	212	6.05	3507
递进关系	7	11.86	15	25.42	19	32.20	9	15.25	9	15.25	59
对比关系	19	31.67	18	30.00	15	25.00	8	13.33	0	0.00	60
假设关系	11	15.94	30	43.48	12	17.39	12	17.39	4	5.80	69
解说关系	434	47.85	260	28.67	130	14.33	50	5.51	33	3.64	907
例证关系	133	52.57	63	24.90	35	13.83	10	3.95	12	4.74	253
目的关系	102	30.63	117	35.14	60	18.02	45	13.51	9	2.70	333
评价关系	82	37.10	66	29.86	51	23.08	14	6.33	8	3.62	221
让步关系	5	31.25	4	25.00	5	31.25	0	0.00	2	12.50	16
顺承关系	186	36.12	155	30.10	107	20.78	51	9.90	16	3.11	515
条件关系	7	9.86	24	33.80	23	32.39	10	14.08	7	9.86	71
推断关系	14	36.84	8	21.05	11	28.95	4	10.53	1	2.63	38
选择关系	2	20.00	4	40.00	1	10.00	1	10.00	2	20.00	10
因果关系	232	33.82	244	35.57	121	17.64	67	9.77	22	3.21	686
转折关系	72	36.73	72	36.73	29	14.80	13	6.63	10	5.10	196
总分关系	51	21.70	82	34.89	67	28.51	30	12.77	5	2.13	235
合计	2342	32.04	2372	32.45	1533	20.97	711	9.73	352	4.82	7310

CDTB 中,层次最大为 9。表 2.8 中,CDTB 中层次为 1 的关系有 2342 个,占有关系的 32.0%;层次为 2 的 2373 个,占 32.4%;层次为 3 的 1533 个,占 21%;层次为 4 的 712 个, 9.7%;其它层有 352 个,前 4 层占 95.18%,说明 CDTB 中篇章结构树总体不是太复杂。对每种关系而言,背景关系(73.1%)、解说关系(47.9%)和例证关系(52.6%)集中在第 1 层,选择关系、目的关系和假设关系多出现在第 2 层。

对 CDTB 中的树结构进行统计，发现某些篇章模式（段落的篇章结构树）出现次数较多，具体统计见表 2.9。表 2.9 中大写字母 A、B、C 和 D 代表子句，用图画出相应的子句结构图，频次表示相应结构在语料中出现的次数。

表2.9 CDTB 中出现较多的篇章模式

编号	结构	频次	编号	结构	频次
1	并列类 	308	2	解说类 	130
3	因果类 	124	4	解说类 	73
5	并列类 	71	6	并列类 	47
7	并列类 	46	8	因果类 	38
9	因果类 	36	10	并列类 	31
22	解说类 	21	12	并列类 	20
13	并列类 	17	14	并列类 	15

由表 2.9 可知，部分篇章结构模式并不太复杂，甚至是有规律可循的，如果不考虑关系，1、2 和 3 的结构一致，4、5、8 和 14 的结构也一致，这为本文后面篇章结构树构建提供参考。观察这些篇章模式，发现除了 2 和 3，剩下的都包含并列类，这也说明了并列类所占比例较大。并列类只有编号 7 包含 3 个篇章单位，其余均包含 2 个篇章单位，由于 CDTB 中包含多个篇章单位的关系规模较小，所以本文第 4 章在做自动分析时将包含多个篇章单位的关系转换成向左二叉树是可行的。

3.5.4 篇章单位主次统计与分析

本文在 2.5 节汉语篇章单位主次定义中提到单中心和多中心,单中心指某一关系中能概括它所在关系层主旨的是该关系的一个篇章单位。多中心指能概括该关系层主旨的是该关系的两个或多个篇章单位。CDTB 中每一种类别的篇章关系的单中心与多中心的统计如表 2.10 所示。

表2.10 CDTB 中单中心与多中心统计表

关系类		单中心		多中心		总数
		个数	百分比(%)	个数	百分比(%)	
因果类	因果关系	673	98.10	13	1.90	686
	推断关系	36	94.74	2	5.26	38
	假设关系	58	84.06	11	15.94	69
	目的关系	330	99.10	3	0.90	333
	条件关系	60	84.51	11	15.49	71
	背景关系	134	100.0	0	0.00	134
因果类小计		1291	96.99	40	3.01	1331
转折类	转折关系	187	95.41	9	4.59	196
	让步关系	15	93.75	1	6.25	16
转折类小计		202	95.28	10	4.72	212
并列类	并列关系	288	8.21	3219	91.79	3507
	顺承关系	119	23.11	396	76.89	515
	递进关系	9	15.25	50	84.75	59
	选择关系	1	10.00	9	90.00	10
	对比关系	51	85.00	9	15.00	60
并列类小计		468	11.27	3683	88.73	4151
解说类	解说关系	887	97.79	20	2.21	907
	总分关系	233	99.15	2	0.85	235
	例证关系	253	100.00	0	0.00	253
	评价关系	221	100.00	0	0.00	221
	解说类小计	1594	98.64	22	1.36	1616
合计		3555	48.63	3755	51.37	7310

从表 2.10 中可以看出背景关系、例证关系和评价关系都是单中心的。而选择关系、顺承关系、对比关系一般含有多个中心。四大类中,并列类多数情况(88.73%)下是多中心的,解说类、因果类和转折类多是单中心的。

对于单中心的关系,中心分在前和在后,表 2.11 是 CDTB 的中心位置统计表。

表2.11 CDTB 中篇章单位主次前后位置统计表

	位置	前中心		后中心		总数
	关系类别	个数	百分比(%)	个数	百分比(%)	
因果类	因果关系	212	31.50	461	68.50	673
	推断关系	18	50.00	18	50.00	36
	假设关系	2	3.45	56	96.55	58
	目的关系	172	52.12	158	47.88	330
	条件关系	5	8.33	55	91.67	60
	背景关系	7	5.22	127	94.78	134
因果类小计		416	32.22	875	67.78	1291
转折类	转折关系	11	5.88	176	94.12	187
	让步关系	0	0.00	15	100.00	15
转折类小计		11	5.45	191	94.55	202
并列类	并列关系	254	88.19	34	11.81	288
	顺承关系	17	14.29	102	85.71	119
	递进关系	1	11.11	8	88.89	9
	选择关系	0	0.00	1	100.00	1
	对比关系	12	23.53	39	76.47	51
并列类小计		284	60.68	184	39.32	468
解说类	解说关系	766	86.36	121	13.64	887
	总分关系	201	86.27	32	13.73	233
	例证关系	249	98.42	4	1.58	253
	评价关系	181	81.90	40	18.10	221
解说类小计		1397	87.64	197	12.36	1594
合计		2108	59.30	1447	40.70	3555

由表 2.11 可知,假设关系、条件关系、背景关系、让步关系、顺承关系、递进

关系、选择关系、对比关系侧重于后中心。而并列关系、解说关系、总分关系、例证关系和评价关系一般中心在前。推断关系、目的关系前后中心比例相当。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。

3.6 本章小结

本章主要介绍基于连接依存树的汉语篇章结构语料库（CDTB）构建。为便于语料标注操作，提出了比较符合汉语篇章结构认知习惯的自顶向下的语料库标注策略。为保证语料标注质量，采用人机结合的语料库标注方法。选取部分语料进行了标注一致性测试，测试结果表明本章所标语料质量达到实用水平。最后从连接词、篇章关系、篇章结构、篇章单位主次等方面对 CDTB 进行了系统的统计分析，统计结果表明 CDTB 规模达到一定程度，统计数据反映的情况和汉语特点基本一致。

本章部分内容整理发表在自然语言处理顶级国际会议 EMNLP2014 上。

第4章 基于 CDTB 的汉语篇章结构分析

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比较缓慢。英语篇章结构分析目前研究较多,而汉语篇章结构分析研究由于语料缺乏,目前主要是对篇章中的某个特定问题进行研究,还没有完整的汉语篇章结构分析平台。因此,本章在第3章构建的 CDTB 语料的基础上,进行系统化的汉语篇章结构分析研究。

4.1 引言

目前篇章结构分析系统研究主要针对英语,比较有代表性有 Soricut 等^[19]、Hernault 等^[37]、Lin 等^[55]、Feng 等^[38]、Joty 等^[39-40]、徐凡^[47]等工作。由于汉语语料缺乏,目前没有完整的汉语篇章结构分析平台。汉语篇章结构分析平台主要包括子句识别、连接词识别与分类、篇章关系识别、篇章单位主次识别等工作,目前没有看到针对汉语篇章单位主次的识别研究工作,其它相关的汉语子句识别、连接词识别与分类和篇章关系识别的研究工作也不多,下面简单介绍目前的研究进展及存在的问题。

子句识别方面,由 2.1 的子句定义可知,标点符号是子句的重要形式标志,根据某些标点(如句号)可直接判定其所分隔的语言片段为子句,而另外一些标点(如顿号)所分隔的语言片段则不可能是子句,还有的标点(如逗号)分隔的语言片段有些情况下是子句,有时情况下不是子句。在自然语言处理的不同任务中都能看到引入标点的研究,有效识别标点功能,有助于句法分析、篇章分析、机器翻译等自然语言处理技术性能的提高。自动识别标点是否为子句边界的研究较少,文献[74]和[75]识别的是句子,在句法树中有明显的 IP 标示,本文所识别的是子句(根据定义可知至少包含 IP 或 VP,主要从语义上进行定义),粒度要细的多。综合以上分析可知,标点对子句分割意义重大,其中逗号尤为关键。

连接词识别与分类方面,包含连接词标记的语料库目前主要有汉语复句语料库、清华汉语树库、哈工大中文篇章关系语料库。胡金柱等^[9294]在汉语复句语料库进行了关系词识别的相关研究,关系词提取的正确率达到 89.8%,连用关系标记标识正确率达 72.9%。洪鹿平^[95]在清华汉语树库上做汉语复句关系自动判断研究,作者穷尽式地收集关系词语,并把关系词语标注上联合和偏正两种类型,然后抽取特征利用 CRF

模型进行分类。王东波等^[96]基于条件随机场进行有标记联合结构的自动识别,使用北京大学的《人民日报》语料和清华汉语树库,分别用基于复杂特征的特征模板和增加语言学特征的特征模板在含有嵌套的联合结构、无嵌套联合结构和最长联合结构语料上进行了实验,F1 值分别为 88.21%、87.85%和 84.42%。李艳翠等^[97]利用规则从清华汉语树库中提取复句关系词并标注其类别,然后抽取自动句法树和标准句法树的句法、词法、位置特征进行复句关系词的识别和分类,实验结果表明复句关系词判断正确率达 95.7%,复句关系词类别判断 F1 值为 77.2%。张牧宇等^[76]在哈工大中文篇章关系语料库上进行句间语义关系的识别。针对显式篇章句间关系,提出基于关联词规则的方法进行识别,取得了很好的效果。综上可知,目前汉语语料多关注句内关系,没有标注句间关系,连接词并不是真正的篇章连接词。目前连接词标注和关心分类多参考英语的分类方法,标注出的连接词种类过多,关系类别和汉语分类体系相差较大。因此本文将采用第3章自建的符合汉语特点的 CDTB 语料进行连接词的识别与分类。

隐式关系识别是指没有连接词的情况下判断两个论元之间是否存在何种逻辑语义关系,由于连接词的缺失,通常只能根据一些语言学特征进行关系的识别,一般正确率不高。汉语方面,Huang 等^[66]进行了汉语篇章关系识别,在自标的四大类关系语料上,利用词、词性、上位词等特征训练分类器。分四大类关系的最好结果时正确率为 88.28%,F1 值为 63.69%。虽然识别效果较好,但识别是隐式显式一起进行的,故不能体现隐式关系的识别情况。张牧宇等^[76]进行中文篇章级句间语义关系识别,针对隐式篇章关系,他们抽取词汇、句法、语义等特征,采用有指导模型进行识别,实验结果表明除“扩展关系”外,其它类型存在高准确率、低召回率的特性。以最大熵模型下的“因果关系”为例,识别准确率达到 68.75%,召回率却只有 8.03%。可见,汉语隐式篇章关系识别任务比较困难。

综上可知,目前汉语篇章结构分析的各个子任务研究均不多,特别是隐式篇章结构分析相当困难。本章重点介绍基于 CDTB 的汉语的篇章结构分析平台。首先给出了汉语篇章结构分析平台框架,该平台采用自底向上的篇章结构树构建方法,综合了子句识别、连接词识别与分类、篇章关系识别、篇章单位主次识别等子任务。然后给出了实验所用特征和实验设置情况。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。

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）。

训练完成后，对测试文本按照图 4.2 所示的流程进行篇章结构树的构建。

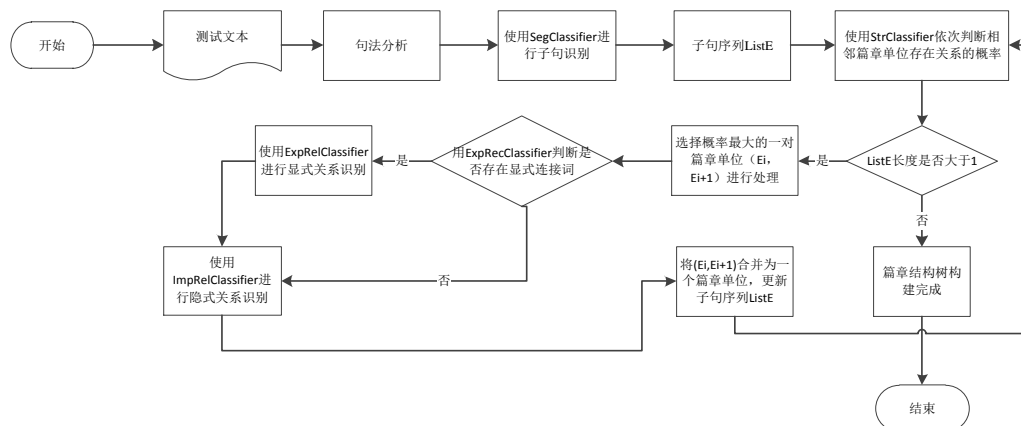


图4.2 汉语篇章结构树构建流程图

如图 4.2 所示，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。然后利用 **StrClassifier** 判断子句之间是否存在关系，判断是否存在关系时分别计算相邻篇章单位存在关系的概率，选择概率最大的一个进行合并，并判断相邻篇章单位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。以上操作循环进行，直到所有子句处理完毕即得到最终的篇章结构树。

本分析平台最重要的步骤是篇章结构树构建，本文采用自底向上的篇章结构树构建方法，算法伪代码如图 4.3 所示。图 4.3 所示的算法中，输入是切分好的子句，输出是篇章结构树。 E_i 表示 $ListE$ 中的第 i 个元素。**Label** 函数是判断 E_i 和 E_{i+1} 之间关系的具体类别，用 **ExpRecClassifier** 判断是否有显式连接词，如果有显式连接词则用显式关系判断分类器 **ExpRelClassifier**，否则用隐式关系判断分类器 **ImpRelClassifier**。**Center** 函数是根据给定主次判断分类器 **CenClassifier** 判断 E_i 和 E_{i+1} 的中心指向。**CreatTree** 函数是创建包含关系和主次信息的子树，子树的左孩子是 E_i ，右孩子是 E_{i+1} 。本章的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。


```

输入: ListE=[E1,E2,...], 文本的子句序列
输出: 篇章结构树 FinalTree;
For Ei in ListE:
    For span(Ei,Ei+1) in ListE:
        # 计算 ei 和 ei+1 两个篇章单位之间存在关系的得分
        Score[i] = Prob(StrClassifier(Ei,Ei+1))
    While len(ListE) > 1:
        #取得分最大的合并
        i = argmax(Score)
        #判断关系
        newLabel = Label(ExpRelClassifier, ImpRelClassifier,
                        ExpRecClassifier, ei, ei+1)
        #判断主次
        newCenter = Center(CenClassifier, Ei, Ei+1)
        #创建新的子树
        newSubTree = CreatTree(Ei,Ei+1,newLabel,newCenter)
        delete(score[i])
        delete(score[i+1])
        #将能合并的单位进行合并
        ListE = [E0,...,Ei-1,newSubTree,Ei+2,...]
FinalTree = E0
Return FinalTree

```

图4.3 自底向上的篇章结构树生成算法

4.3 实验方法

4.3.1 所用特征

本章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 4.2 可知,本分析平台可以拆分成多个子任务,每个子任务目标不一样,所用的特征也有所差别,下面详细介绍本平台中各个子任务所用的特征。

4.3.1.1 子句识别

子句识别任务主要参考 Xue 等^[74]的特征, 加上针对本文汉语子句识别任务简单特征, 抽取每个标点的特征进行实验, 所用特征及说明见表 4.1。

表4.1 子句识别所用特征及说明表

特征组号	特征	说明
1	F1_P_N F1_W_N	从逗号到前一逗号或句首的范围内前面 N 个词的词性及词, 后面 N 个词的词性及词
	F2_P_N	
	F2_W_N	
2	F3 F4	逗号之后第一个词的词性和词
3	F5_1 F5_2	逗号前后单元所包含的属于连接词表中的词
4	F6 F7 F8	逗号左右兄弟及其组合的句法信息标记
5	F9	逗号左右兄弟及逗号父亲句法信息组合
6	F10_1 F10_2 F10_3	从逗号到前一逗号或句首是否包含是 (VC)、表语形容词 (VA)、有 (VE)、其它动词 (VV), 从属连词 (CS); 从逗号到后一逗号或句尾是否包含是 (VC)、表语形容词 (VA)、有 (VE)、其它动词 (VV), 从属连词 (CS);
	F10_4 F10_5 F11_1	
	F11_2 F11_3 F11_4	
	F11_5	
7	F12 F13 F14	逗号的父节点是否为并列的 IP 结构、逗号是否为第一层子节点、逗号父节点是否是第一层并列 IP 结构
8	F15	逗号所在句子标点的集合
9	F16 F17	逗号到前一标点或句首句子长度是否小于 5, 逗号左右两边句子长度差是否大于 7
10	F18	逗号所在句法树层次
11	F19 F20 F21	逗号的父节点、左兄弟节点、右兄弟节点是否为 NP
12	F22 F23	从本标点 to 前一个标点或句首的句子中, 第一个词及最后一个词的词性及词的组合
13	F24	逗号前后单元包含的相同词及词性信息

例 2.1.4 中第 1 句话共有两个逗号, 根据表 4.1 所述特征, 抽取第 1 个逗号和第 2 个逗号相应特征值实例, 一行表示一个逗号的特征实例。实例中第 1 行表示逗号在整个语料中的顺序号, 第 2 行表示逗号是否为子句边界, +1 表示标点为子句边界, -1 表示非子句边界。抽取逗号特征实例时用到第 1 句话的句法信息, 相应的句法树如图 4.4 所示:

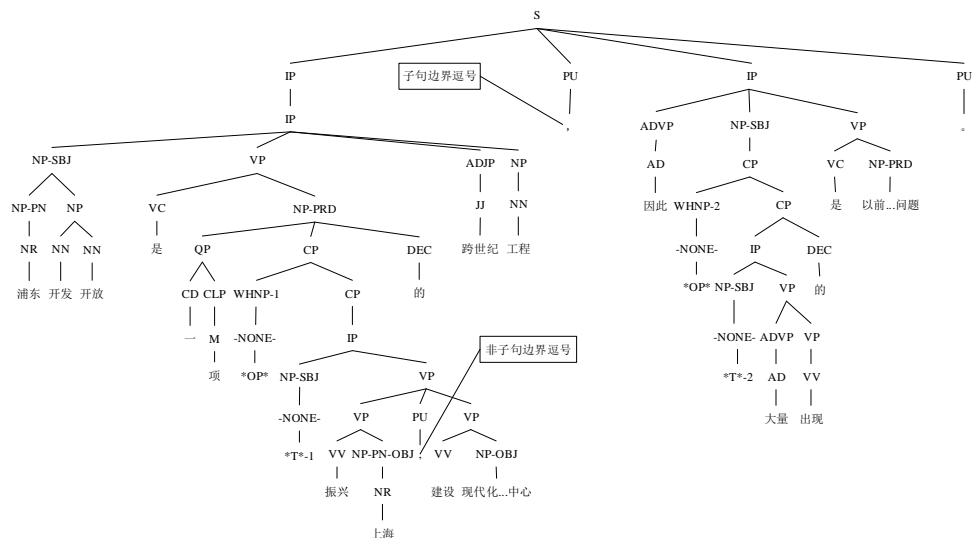


图4.4 例 2.1.4 中第 1 个句子句法树信息

例 2.1.4 中第 1 个和第 2 个逗号最终得到的特征实例如例 4.1 和例 4.2 所示。

例4.1 -1 F1_P_1=NR F1_W_1=浦东 F1_P_2=NN F1_W_2=开发 F1_P_3=NN
F1_W_3=开放 F2_P_1=NR F2_W_1=上海 F2_P_2=VV F2_W_2=振兴 F2_P_3=M
F2_W_3=项 F3=VP F4=建设 F6=VP F7=VP F8=VP+VP F9=VP+VP+VP F10_1=yesVC
F10_4=yesVV F11_4=yesVV F15=, +, +。 F17>7 F18=8 F19=yesNP F22=NR+NR F23=浦东
+上海

例4.2 +1 F1_P_1=VV F1_W_1=建设 F1_P_2=NN F1_W_2=现代化 F1_P_3=NN
F1_W_3=经济 F2_P_1=NN F2_W_1=工程 F2_P_2=JJ F2_W_2=跨世纪 F2_P_3=DEC
F2_W_3=的 F3=VP F3=AD F4=因此 F5_2=因此 F6=IP F7=IP F8=IP+IP F9=IP+S+IP
F10_1=yesVV f11_1=yesVV f11_2=yesVC F15=, +, +。 F17=1 F22=VV+NN F23=建设+工
程 F24=DEC

4.3.1.2 连接词识别与分类

本文抽取简单的词法、句法和位置特征进行连接词的识别和分类。所用特征如下，括号中内容以例 2.1.4 中的“因此”为例进行说明。

● 词法特征

- 2 连接词前后的 2 个词及词性 (NN 工程 PU , AD 大量 VV 出现)

● 句法特征

- 1 连接词节点句法信息 (ADVP)
- 2 连接词节点父亲节点信息 (IP)
- 3 连接词节点左兄弟信息 (NULL)
- 4 连接词节点右兄弟信息 (NP)

- 位置特征

- 1 连接词是否位于句首（否）
- 2 连接词前是否有标点（是）

4.3.1.3 篇章关系及主次识别

借鉴 Lin 等^[51]在 PDTB 上进行隐式关系识别时所用的特征,结合问题本身的特点,本文提取下面三组特征。

- 上下文特征

CDTB 中,每一段话构成一个篇章结构树,由于本章对篇章结构树先进行了二义化处理,所以每个关系包括两个篇章单位 (Arg1, Arg2)。CDTB 中一个篇章关系可能包括多于一个的关系类别,如例 4.3 所示,这种情况下,我们在生成实例时将它表示成两个具有不同类别的关系。

例4.3 所有境内机构借用国际商业贷款应当经国家外汇管理局批准。|未经批准而擅自签订的国际商业贷款协议无效,外汇局将不予办理外债登记,银行不得为其开立外债专用帐户,借款本息不得擅自汇出。(转折类、并列类) (chtb_0119)

观察处理后的语料可以发现篇章关系对之间存在完全嵌入论元和共享论元两种最普遍的模式,如图 4.5 所示。

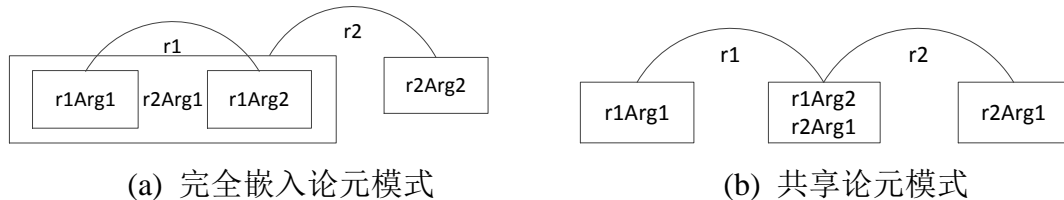


图4.5 两种普遍篇章模式

借鉴 Lin 的上下文特征思想,根据 CDTB 中的具体情况,给出了完全嵌入论元和共享论元两种篇章模式的六个特征,如下所示。curr 指当前关系,pre 指上一个关系,next 指下一个关系。其中完全嵌入论元模式与 Lin^[51]的不同,本文判断当前关系是否完全嵌入上一个关系的 Arg1 或者 Arg2 中,下一个关系是否完全嵌入到当前关系的 Arg1 或者 Arg2 中。当上一个关系或者下一个关系为显式关系时,用它们的连接词作为上下文特征,此特征记为 FContext。

完全嵌入论元模式:

curr embedded in pre.Arg1
curr embedded in pre.Arg2

共享论元模式:

prev.Arg2=curr.Arg1
curr.Arg2=next.Arg1

next embedded in curr.Arg1

next embedded in curr.Arg2

● 词汇特征

1 词对特征：在英语篇章关系处理中，已有实验证明，词对特征在关系的识别中非常有效。特征 FWP(w1,w2)，w1 指 Arg1 中一个词，w2 指 Arg2 中一个词。

2 词和词性：句子中的动词在一定程度上能够反应出句子的意图，因此判断论元是否有以下的词性标注："VV"、"VC"、"VE"、"VA"、"CS"、"CC"、"AD"、"DEV"、"BA"、"SB"、"LC"，如有则给出该词性对应的词和词性及其组合。此特征记为 FVwp。

● 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。借鉴 Lin 等^[51]的特征，在本文中，首先，我们利用 Stanford 句法分析器对每个句子进行依存句法分析，然后从每个论元对应的依存树中选择所有拥有被支配者的词和依存类型。每个论元根据其跨度的不同可能对应着完整依存树，子树或者多棵树。

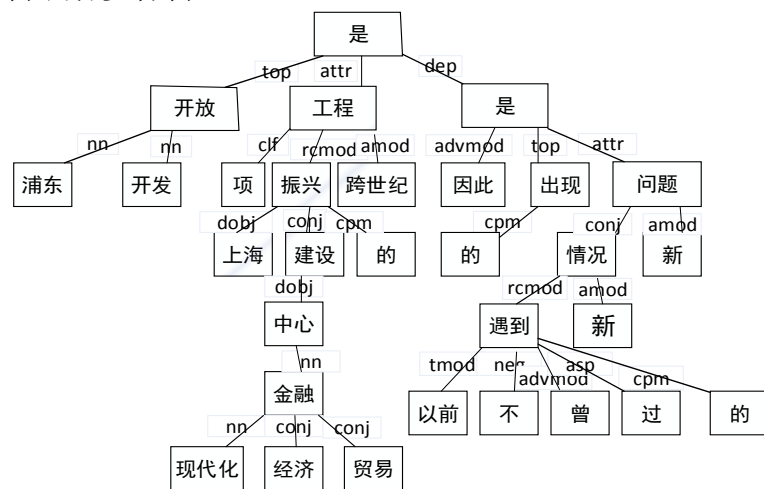


图4.6 例 2.1.4 中第 1 句的依存分析树

图 4.6 给出了例 2.1.4 中第 1 句的完整依存分析树，对于整棵树我们可以抽取如下 的 依 存 规 则：是_top_attr_dep，开 放_nn_nn，工 程_clf_rcmod_amod，是_advmod_top_attr，振 兴_dobj_conj_cpm，以此类推，遍历整个依存树。核对每一个规则是否出现在 Arg1 中，Arg2 中和同时出现在两者中，最后表示为三个二元特征。此特征记为 FDep。

4.3.1.4 篇章结构识别

本文采用 Feng 等^[38]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，所以他们提到的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了以下特征：

- 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 Arg1 和 Arg2 中的词对是否在同一个中类中，如果在，就抽取这个词对及其词性。

- 上下文信息

- 1 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
- 2 为连接词及其前后词的词性
- 3 同 4.3.1.3 节所述上下文特征

- 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 4.3.1.3 节所述依存树特征。

4.3.2 实验设置

CDTB 共有 500 个文档（chtb001-chtb0657），在 CTB 中句子编号从 1 到 6648。每个篇章（段落）标注为一棵连接依存树，共有效标注 2342 个篇章。本章实验统一取 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 4.2 给出 CDTB 中训练集和测试集的划分及标注情况，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 4.2 中给出的是语料中实际标注的情况，在没有转化为二叉树前，一个篇章关系可以包含多个子句。

表4.2 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 3372	单中心 311 (中心在前 207, 中心在后 104), 多中心 383

篇章结构分析的各项工作中,子句(基本篇章单位)识别是一项基础工作,所谓子句识别就是自动对输入文本进行篇章单位分割,只有识别了子句,才能在此基础上做进一步的篇章结构分析。CDTB中共有15985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10937个,非子句边界的标点5048,正例占68.4%,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。CDTB中共有逗号9552个,其中标注为子句边界的逗号有5304个,标注为非子句边界的逗号有4248个,正例占55.5%。子句考虑逗号的情况共抽取8698个训练样例(正例4780个,负例3918个),854个测试样例(正例524个,负例330个);子句识别只考虑句内标点共抽取8872个训练样例(正例4797个,负例4075个),878个测试样例(正例536个,负例342个)。分别使用NLTK⁵工具包中的决策树、最大熵、贝叶斯进行实验。

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

⁵ <http://www.nltk.org/>

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，如“和”有些时候是篇章连接词，为正例，但大多数情况下为负例。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，如原来的“不但……而且”、“不但”、“而且”，经处理后只剩“不但”和“而且”。实验共抽取 10923 个实例，其中训练样例 10016 个（其中是连接词的实例有 1935 个，非连接词实例为 8081 个），测试样例 907 个（其中是连接词的实例有 186 个，非连接词实例为 721 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（解说类 201，转折类 237，因果类 427，并列类 1073），测试样例 186 个（解说类 11，转折类 18，因果类 44，并列类 114）。

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（解说类 1275 个，转折类 40，因果类 788，并列类 3588），测试实例 617 个（解说类 140，转折类 1，因果类 79，并列类 397）。出现多个关系类别时，扩展实例使之变为多个单关系类别。

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 4.1 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，子句 b 由于只是第 1 句的一部分，因此也不和第 2 句 cdefg 的组合产生负例，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。实验对输入是已经标注好的子句和本文自动分割的子句，输出是二叉树和多叉树的结构和关系分别进行评价。

对于篇章结构的打分，本文采用标准的 Parseval^[98]打分矩阵，也就是正确的篇章结构占分析得到的所有篇章结构的百分比作为准确率 P，正确分析的篇章结构占参考答案所有篇章结构的比值作为召回率 R，以及综合考虑准确率 P 和召回率 R 的打分 F1 值。

4.4 实验结果及分析

4.4.1 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,剩下的可能为子句边界的冒号、破折号、逗号等为句内标点。若无特殊说明,本节实验均指句内标点识别,句内标点子句识别整体实验结果如表 4.3 所示。为验证本文所提特征的效果,文本将 Xue 等^[74]的工作重现,由于 Xue 等人所做的工作是自动识别起句号作用的逗号,和本文的任务不尽相同,实验数据不好直接对比,故本文仅使用 Xue 等人的特征进行本文的子句识别,没有将其论文中的数据与本文数据直接进行对比。本文记标点为子句边界的情况为正例,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 4.3。

表4.3 句内标点是否为子句边界识别结果

分 类 器	使用本文特征		使用文献[74]特征	
	标准句法树	自动句法树	标准句法树	自动句法树
	Acc. F1(+) F1(-)	Acc. F1(+) F1(-)	Acc. F1(+) F1(-)	Acc. F1(+) F1(-)
最大熵	93.9 95.0 92.3	91.8 93.2 89.6	93.5 94.7 91.8	90.9 92.4 88.6
决策树	74.2 81.9 55.1	73.1 81.3 51.5	92.7 93.9 90.8	82.1 85.7 76.2
贝叶斯	91.6 93.0 89.7	88.8 90.6 86.3	91.6 93.0 89.6	88.8 90.7 85.8

从表 4.3 可以看出,使用本文的特征要比使用 Xue 等人的特征结果好,说明本文所用的特征还是十分有效的。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 4.3 可以发现,F1(-)比 F1(+)效果要差,即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。对于逗号是否为子句边界的实验结果如表 4.4 所示。

表4.4 逗号子句边界识别结果

分类器	标准句法树			自动句法树		
	正确率	F1(+)	F1(-)	正确率	F1(+)	F1(-)
最大熵	93.8	95.1	92.3	91.2	92.8	88.6
决策树	74.5	82.3	54.0	73.3	81.7	50.7
贝叶斯	92.1	93.5	89.9	89.1	91.0	86.1

由表 4.4 可知, 上述三个分类器中最大熵的效果最好, 采用 CTB6.0 中所提供的标准句法树, 最好的正确率为 93.8%, 采用 Berkeley 自动句法分析树, 正确率是 91.2%。对于逗号为子句边界的情况, 使用标准句法树 F1 值最高 95.1%, 自动句法树 F1 值最高 92.8%。对于逗号非子句边界的情况, 使用标准句法树 F1 值最高 92.3%, 自动句法树 F1 值 88.6%。从表 4.4 可以看出, 逗号非子句边界的情况比逗号为子句边界的情况识别效果差, 主要归纳原因如下: 训练样例不均衡, 正负比大约为 1.3:1; 逗号是非子句边界的情况比较复杂, 如有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, 判断起来比较困难, 且较难找到非常有效的特征。

实验中约有 10% 的情况子句识别错误, 下面对其错误情况进行分析。

1) 负例被错误地识别为正例

从上面实验可以看出, 负例识别的效果较差, 被错误识别的逗号有以下情况: 子句内部主语与谓语之间的逗号; 子句内部动词与宾语之间的逗号; 子句内部状语之间的逗号。被错误识别的逗号用下划线标示, 具体情况如下:

例4.4 出口快速增长,成为推动经济增长的重要力量。 (chtb_0097)

例4.5 确立了“以资源换技术,以产权换资金,以市场换项目,以存量换增量”的利用外资新思路。 (chtb_0091)

例4.6 天津港保税区投入运行五年来,已建成了中国第一货物分拨中心, 具备了口岸关的功能, 开通了天津港保税区经西安、兰州到新疆阿拉山口口岸的铁路专用线。 (chtb_0099)

例 4.4 中的逗号前面是整个句子的主语, 逗号不能作为子句边界, 但句法分析将“出口快速增长”分析为 IP, 导致此逗号没有正确分类; 只使用词法信息, 其中包含出口 (NN) 和增长 (VV), 也较易识别为正例。例 4.5 中的逗号是句内动词和宾语之间的逗号, 但系统识别时将其识别为子句, 因为现在没有体现他们是句子宾语的信息。例 4.6 中的“天津港保税区投入运行五年来”是整个句子的时间状语, 逗号表示句子内部状语之间的停顿, 但“天津港保税区投入运行五年”是个子句, 从而导致识别错误。

2) 正例被错误地识别为负例

例4.7 内地经济长期稳定地增长,香港经济将从充满活力的内地经济中获益。 (chtb_0093)

例 4.7 中的逗号前的“增长是”动词, 逗号后的“香港”是名词, 导致分类器没有正确分类。

4.4.2 连接词识别与分类

本文主要进行显式连接词的自动识别与分类实验,实验与参考文献[97]基本相同,但文献[97]主要是对清华汉语树库中的连接词进行识别与分类,而且清华汉语树库中并没有准确标注连接词信息,也没有句子与句子之间的连接词。本文所用 CDTB 是以段落为单位进行标注的,连接词不仅有连接子句的连接词,还有连接句子的连接词。

4.4.2.1 连接词识别

连接词识别所用的特征如 4.3.1.2 所示,主要有词汇、句法和位置特征,表 4.5 给出了连接词识别的各个特征及特征组合的结果。

表4.5 是否为连接词识别正确率

语料库	特征	自动句法树			标准句法树		
		最大熵	决策树	贝叶斯	最大熵	决策树	贝叶斯
本文 CDTB	词汇	86.2	87.3	81.5	86.7	87.2	81.8
	句法	80.9	82.2	80.3	84.3	84.7	82.4
	词汇+句法	86.6	88.1	81.9	87.9	88.9	83.7
	词汇+句法+位置	87.2	88.4	83.4	88.2	88.5	85.3
清华树库	词汇+句法+位置	91.2	92.1	88.1	-	-	-

从表 4.5 可以看出,使用决策树效果最好,说明连接词识别问题并不太复杂,可以抽取出一一定的规则。由实验可知,单纯的词汇特征对连接词的识别也有一定作用。利用词汇、句法和位置特征的组合进行连接词的识别效果最好,使用自动句法树和标准句法树连接词识别的正确率分别为 88.4%和 88.5%。本文实验所用特征和文献[97]相同,表 4.5 给出文献[97]在清华树库上的部分实验结果。从表 4.5 可以发现,使用决策树,本文实验结果比文献[97]低 3.7%,主要原因是本文语料采用自标的 CDTB,连接词包括句内连接词和句间连接词,而文献[97]只考虑句内连接词,连接词由算法抽取,抽取时只考虑出现次数最多的连词(c)、副词(d)和连接词(l)三种词性,而本文没有做此限制,使用的连接词范围较广。

表 4.5 中给出的实验结果包含需要识别的词不是连接词的情况,这类词所占比例较高(占 80%),故而总正确率较高。表 4.6 给出利用决策树和最大熵分类器,使用词汇、句法和位置特征对是连接词的词进行识别的准确率、召回率和 F1 值。从表 4.6 可知,对于连接词的识别,利用最大熵分类器的效果明显要好。各种情况下连接词的

识别准确率均高于召回率。使用自动句法树和标准句法树结果相差较小,使用自动句法树的 F1 值为 69.2%,使用标准句法树 F1 值 69.3%,说明连接词识别任务对句法分析性能的好坏依赖程度较小。

表4.6 连接词识别的准确率、召回率和 F1 值

分类器	类别	准确率	召回率	F1 值
最大熵	自动句法树	78.8	61.8	69.2
	标准句法树	78.9	61.8	69.3
决策树	自动句法树	56.8	49.6	52.3
	标准句法树	58.9	48.5	52.7

4.4.2.2 连接词分类

连接词识别完成后,需要对连接词进行分类,本实验使用连接词识别时所用的特征,利用最大熵分类器,分别对给定连接词和自动识别连接词进行分类实验。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如表 4.7 所示。

从表 4.7 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“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。

表4.7 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

通常,我们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,因此首先需要确定某个词是否为连接词,然后对识别为连接词的词进行分类。实验分别抽取是否为连接词和连接词类别数据,测试时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 4.8 给出在连接词自动识别基础上进行连接词分类的每个类别准确率、召回率和 F1 值。仅从数据上看因果类和转折

类 F1 值高于给定连接词的结果,但二者之间没有可比性,表 4.8 是在连接词自动识别正确的情况下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。

表4.8 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

分析实验结果,我们发现大部分篇章连接词只表示一种关系,有 34 个篇章连接词(占有所有连接词的 12.4%)可以表示多种关系,识别错误主要由这部分连接词的歧义导致。这类连接词总数虽然不多,但包含的常用连接词较多(例如并、而、但)。同一连接词对应的关系类型越少、类型越集中,该词的歧义性越小。以显式连接词“而”为例,它对应的具体关系类型及关系大类分别如下:并列类 65 次(其中并列关系 39 次,对比关系 19 次,递进关系 6 次,顺承关系 1 次);转折类 4 次(转折关系 4 次);因果类 1 次(因果关系 1 次)。因此,连接词识别与分类的主要难点为是否为连接词的歧义及连接词关系类别的歧义问题。

4.4.3 隐式篇章关系识别

篇章结构分析的另一个关键内容是篇章关系的识别,本节实验主要识别隐式的 4 大类篇章关系。由 4.3.2 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 4.3.1.4 所述特征和最大熵分类器,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。每种类别具体的准确率、召回率和 F1 值如表 4.9 所示。

隐式篇章关系识别任务中,隐式篇章关系共有 6305 个,训练实例 5688 个(解说类 1276 个,转折类 38,因果类 786,并列类 3588),测试实例 617 个(解说类 140,转折类 1,因果类 79,并列类 397)。

表4.9 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 4.9 给出了四大类关系识别的具体准确率、召回率和 F1 值，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，所以并列类的识别最优。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。由实验设置可知，解说实例有 1276 个，占据 22.4%，因果实例有 786 个，占据 13.8%。但是转折实例只有 38 个，仅为 0.7%，在测试实例中，共有关系实例 617 个，转折类 1 个，仅占 0.2%，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。从表 4.9 中可以看出，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 4.10）。去除转折类后总正确率为 67.2%，比四大类结果提升 0.3%。从表 4.10 中可以看出，因果类和解说类关系识别的 F1 值均有所提升，因果类提升了 7.5%，解说类提升了 2.6%，但是并列类的 F1 值却下降 0.8%。这种变化主要是由于数据类别的减少，去除转折类后，原来的 4 元分类变为 3 元分类，导致结果发生变化。

表4.10 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

张等^[76]在哈工大篇章结构语料 HIT-CDTB 上进行了隐式篇章关系识别的尝试，但由于 HIT-CDTB 采用 PDTB 系统进行标注，关系分类采用六大类，跟本文所提的连接依存树理论体系不一样，因此不好直接对比。分析张等^[76]的结果发现他们有并列类和因果类两个关系类别，因果类识别准确率 68.7%，召回率 8.0%，F1 值 14.4%，并列类准确率 32.3%，召回率 31.8%，F1 值 32.1%，这两种类别的实验结果单从数据

上比较明显低于本文的结果。

4.4.4 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。各种类别准确率、召回率和 F1 值如表 4.11 所示。

从表 4.11 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。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%。

表4.11 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

4.4.5 基于连接依存树的汉语篇章结构分析平台性能

汉语篇章结构分析平台包括子句识别、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。该平台可以对输入生文本进行处理,生成汉语篇章结构树。

篇章结构分析平台采用 4.2 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,采用图 4.3 所示的自底向上的方法进行汉语篇章结构树的构建。篇章结构树的构建非常关键,本节首先分别给出结构、关系的识别结果,然后给出系统整体性能。

4.4.5.1 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，用来构成无标签的篇章结构树，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进行实验，结果如表 4.12 所示。

表4.12 结构识别结果

类别	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

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 4.12 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句内和句间。

比较表 4.12 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[38]中，英语的句内结构的识别结果均要好于句外。一个原因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 2.1.4 中的第 2 句共有 107 个字符，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的篇章单位之间的具体关系类别进行分类。关系包含显式和隐式，关系分类采用 4 大类。实验结果见表 4.13。

由表 4.13 可知，句内关系类别判断正确率为 78.4%，句间 85.4%，综合 76.2%。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。将表 4.13 和表 4.9 对比可知,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的效果要好。

表4.13 4 大类关系识别结果

类别	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				

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以分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。

在是否存在关系判断时,由于某些连接词会干扰判断结果,如例 4.8 中, a 和 b、c、d 的组合是转折关系, a 和 b 不直接发生关系,但 b 中有一个连接词“然而”,导致分类器错误的判断 a 和 b 之间有关系。

例4.8 a 数年前,北海还是北部湾一个默默无闻的小渔村,||b 然而三五年时间北海已建成了一个现代化都市的框架,|||c 街上客流如潮,||||d 楼房拔地而起。(chtb_0008)

在关系类别判断时,由于并列类实例较多,其它关系类别很容易被错判成并列类,如例 4.9 中两句话之间从意义上可以看出是因果类,但没有特别有效的特征识别这种关系,因此这个例子被错判为并列类。

例4.9 a 但他却无法用“跳”来表达自己的激动之情。|b 三岁时一场高烧,使他患上了严重的小儿麻痹后遗症,||c 这一年他被福利院收养。(chtb_0222)

4.4.5.2 篇章结构树构建结果与分析

表 4.14 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 4.14 对四种情况分别进行了实验，标准子句指语料中手工标注的子句，自动识别子句是本文基于句内标点自动识别的子句。

- 1 使用标准子句和标准句法树
- 2 使用标准子句和自动句法树
- 3 使用自动识别子句和标准句法树
- 4 使用自动识别子句和自动句法树

表4.14 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	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	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	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	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 4.14 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如图 4.1 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%，但处理流程开始的错误逐级向下传递。

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。

由于本平台是在 CDTB 上的首次分析尝试,没有相关的实验可以对比。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,和本文类似,目前报告的最好结果是:结构分析的 F1 值为 85.7%^[41],关系分类(18 类)的 F1 值为 65.12%^[38],“结构+主次”70.95%^[4237],“结构+主次+关系”F1 值为 61.75%^[42]。英语篇章结构分析研究较多,特别是子句切分基本接近标准子句,虽然汉英篇章语料、标注内容等数据不能直接对比,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。

4.5 本章小结

本章主要在自建的 CDTB 语料上进行汉语篇章结构自动分析研究,实现了一个自底向上的汉语篇章结构分析平台。该平台包括子句识别、连接词识别与分类、隐式篇章关系识别、篇章单位主次判断等任务,目标是输入生文本,输出包含子句、关系类别、篇章单位主次和篇章层次结构的汉语篇章结构树。

由于子句之间一定有标点分隔,因此本文采用基于标点的子句识别方法,实验结果表明本文子句定义合理,自动分割子句切实可行。对于显式篇章关系,连接词的识别比较关键,本章采用有监督的方法进行连接词的识别与分类,并给出初步实验结果。对于隐式篇章关系,本章进行了 4 大类的关系识别实验。对于篇章单位主次,本章识别了每个关系的左中心、右中心和多中心三种主次分类情况。最后采用自底向上逐级规约的方法构建汉语篇章结构树,并结合以上研究给出了完整的汉语篇章结构分析平台。该平台既可以进行单个任务的实验,也可以进行整体分析,实验结果可以为今后汉语篇章结构分析研究提供参考。

本章子句识别相关内容发表在《北京大学学报》(自然科学版)2013 年第 1 期、《中文信息学报》2014 年第 5 期、《International Journal of Signal Processing, Image Processing and Pattern Recognition》2015 年第 3 期;连接词识别与分类研究内容发表在《北京大学学报》(自然科学版)2015 年第 2 期;隐式篇章关系识别研究内容发表在《北京大学学报》(自然科学版)2014 年第 1 期;基于连接依存树的汉语篇章结构分析平台研究内容被全国第十四届计算语言学会议(CCL 2015)录用。

第5章 总结与展望

随着语料库的普及,在字、词、句子的层面上,基于统计的自然语言处理取得了前所未有的重视和发展。随着研究的深入,需要考虑句子内部及句子之间的语义关系,研究子句或句子如何组成更大的语言单位——篇章。篇章研究的一个重要工作是篇章结构分析,目前篇章结构分析研究主要针对英语,汉语篇章结构理论不完善,进而相应的语料资源缺乏,计算分析研究较少。本文旨在研究汉语篇章结构的理论表示体系,构建汉语篇章结构资源并进行初步的汉语篇章结构分析。

5.1 总结

本文针对汉语篇章结构分析进行研究,研究工作及其所取得的成果可以概括为以下三个主要方面:

1. 基于连接依存树的汉语篇章结构表示体系研究。本文首次综合 RSTDT 和 PDTB 的优点,结合汉语本身特点,提出基于连接依存树的篇章结构表示方法。此表示方法的叶子节点为子句,内部节点为连接词,连接词本身可以表示篇章关系,其层级地位也可以表示篇章层次结构,存在关系的篇章单位之间根据篇章意图有主次之分。通过大量实例说明本体系结构是符合汉语特点的,并且具有较强的可操作性。与相关理论的对比说明,在理论基础上和汉语特点切合性上,本体系均有一定优越性。

2. 基于连接依存树的汉语篇章结构语料库标注。采用本文所提基于连接依存树的汉语篇章结构表示体系,在 CTB6.0 上标注了 500 个文档的汉语篇章结构语料 (CDTB)。根据标注理论,结合汉语特点,采用自顶向下的标注策略和人机结合的标注方法进行 CDTB 语料库构建。标注实验表明语料标注一致性较好,语料库统计表明所标资源达到了一定的规模,所标的各个部分内容均符合汉语特点。

3. 基于 CDTB 的汉语篇章结构分析。本文给出基于 CDTB 的汉语篇章结构分析平台,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,输出是简化了的连接依存树。实验结构表明本文所提基于连接依存树的汉语篇章结构表示体系是合理的,本文基于此表示体系构建的汉语篇章结构语料 CDTB 是可用的。

本文的创新点主要表现在：1) 理论研究上，本文首次提出了基于连接依存树的汉语篇章结构表示体系，用连接词表示篇章结构层次和篇章关系，该理论简单有效且贴合汉语实际；2) 基础资源构建上，在所提基于连接依存树的汉语篇章结构表示体系的指导下，本文构建了一定规模（500 个文档）的汉语篇章结构语料库（CDTB），该语料库构建过程中充分考虑了汉语特点及语料的可计算性，一致性分析表明本语料库达到实用标准。CDTB 既可为语言学研究提供基础数据，又可用来进行汉语篇章结构自动分析研究；3) 计算分析研究上，给出了汉语篇章结构分析平台，该平台包括子句识别、连接词识别与分类、篇章关系识别和汉语篇章单位主次几个任务，平台综合以上任务，采用自底向上的方法进行汉语篇章结构分析，这是汉语篇章结构系统化分析的首次尝试，为后续研究提供基础平台。

本文的主要贡献在于进行了系统化的汉语篇章结构研究。提出了基于连接依存树的汉语篇章结构表示体系，在此体系指导下标注了一定规模的汉语篇章结构语料库，并在语料库上进行汉语篇章结构分析研究。本文的理论可以丰富发展篇章理论，资源可以为后续汉语篇章结构分析研究提供基础数据，实验分析结果可以为汉语篇章结构分析提供实验对照和参考。

5.2 展望

本文的研究虽然在汉语篇章结构分析方面取得了一定的成果，但是离实用化的目标还有很长的路要走，存在着很多需要进一步探索和研究之处。根据目前的研究状况，下一阶段的研究拟从以下几个方面展开：

1. 提高汉语篇章结构分析性能。本文虽然给出了汉语篇章结构分析结果，但整体效果还不尽如人意，需要提出新的方法对计算模型进行改进。一方面拟利用语料库中显式连接词可删除和隐式连接词可添加的信息，提高隐式篇章关系的识别性能；另一方面拟参考句法分析，采用全局优化的方法进行汉语篇章结构分析。

2. 汉英篇章结构平行语料库构建及分析研究。本文所提基于连接依存树的方法经初步实验也适用英语，如果在本文所提理论指导下标注汉英平行篇章结构对齐语料库，则可以进一步验证本文所提理论，平行语料将有助于汉英机器翻译的研究。

3. 汉语篇章结构分析在自然语言处理中的应用研究。汉语篇章结构分析的结果对自动文摘、指代消解等均有帮助，需要研究如何将本文篇章分析结构集成到其它自然语言处理相关应用系统中提高性能。

参考文献

- [1] 宗成庆. 统计自然语言处理 (第 2 版) [M]. 北京: 清华大学出版社, 2013:276
- [2] Schank R. C. and Abelson R. P. Scripts, Plans, Goals and Understanding: An Inquiry Into Human Knowledge Structures[J]. The Artificial Intelligence Series, 1977.
- [3] De Beaugrande R. and Dressler W. Introduction to Text Linguistics[J]. London and New York: Longman Paperback, 1981.
- [4] Renkema J. Discourse Studies[M]. Amsterdam: John Benjamins 1993.
- [5] Halliday M. A. and Hasan R. Language, context, and text: Aspects of language in a social-semiotic perspective[M]. 2nd ed. Oxford: Oxford University Press, 1989.
- [6] Hobbs J. R. Coherence and coreference[J]. Cognitive Science, 1979, 3(1):67-90.
- [7] Hobbs J. R. Information, Intention, and Structure in Discourse: A first draft[C]. In Burning Issues in Discourse, NATO Advanced Research Workshop, 1993:41-66.
- [8] Mann W. C. and Thompson S. A. Relational propositions in discourse[J]. Discourse processes, 1986, 9(1): 57-90.
- [9] Mann W. C. and Thompson S. A. Rhetorical structure theory: A theory of text organization[M]. University of Southern California, Information Sciences Institute, 1987.
- [10] Mann W. C., Matthiessen C., and Thompson S. A. Rhetorical structure theory and text analysis[J]. Discourse description: Diverse linguistic analyses of a fund-raising text, 1992: 39-78.
- [11] Grosz B. J. and Sidner C. L. Attention, intentions, and the structure of discourse[J]. Computational Linguistics, 1986, 12(3):175-204.
- [12] Grosz B. J., Weinstein S. and Joshi A. Centering: A Framework for Modeling the Local Coherence of Discourse[J]. Computational Linguistics, 21(2), 1995.
- [13] De Beaugrande R. and Dressler W. U. Introduction to text linguistics[M]. London and New York: Longman, 1981.
- [14] Prasad R., Dinesh N., et al. The Penn Discourse Treebank 2.0[C]. In Proceedings of

- LREC, 2008:2961-2968.
- [15] PDTB Research Group. The Penn discourse treebank 2.0 annotation manual[R]. IRCS Technical Reports Series, 2007, 99p.
- [16] Halliday M. and Hasan R. Cohesion in English[M]. London:Longman, 1976.
- [17] Martin, J. R. English Text: System and Structure[M]. Amsterdam and Philadelphia: John Benjamins, 1992.
- [18] Grimes, J. The Thread of Discourse[M]. The Hague: Mouton, 1975.
- [19] Soricut R. and Marcu D. Sentence level discourse parsing using syntactic and lexical information[C]. In Proceedings of HLT-NAACL 2003, 2003: 149-156.
- [20] Skadhauge P. R. and Hardt D. Syntactic identification of attribution in the RST Treebank[C]. In Proceedings of the Recent Advances in Natural Language Processing (RANLP 2005), 2005: 57-61.
- [21] Marcu D. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts[D]. PhD Thesis, Department of Computer Science, University of Toronto, 1997.
- [22] Marcu D. The Theory and Practice of Discourse Parsing and Summarization[M]. MIT Press, 2000.
- [23] Marcu D., Carlson L., and Watanabe M. The Automatic Translation of Discourse Structures[C]. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference, 2000:9-17.
- [24] Marcu D. and Echihiabi A. An unsupervised approach to recognizing discourse relations[C]. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002:368-375.
- [25] Moser M. and Moore J. D. Toward a synthesis of two accounts of discourse structure [C]. Computational Linguistics, 22(3), 1996.
- [26] Asher N. and Lascarides A. Logics of Conversation[M]. Cambridge Univ. Press, 2003.
- [27] Lascarides A. and Asher N. Segmented discourse representation theory: Dynamic semantics with discourse structure[C]. In Computing meaning, 2007: 87-124.
- [28] Scha R. and Polanyi L. An augmented context free grammar for discourse[C]. In

- Proceedings of the 12th conference on Computational linguistics, 1988:573 -577.
- [29] Gardent C. Discourse tree adjoining grammars[R]. Technical report, University of Saarland, CLAUS report, 1997, 38p.
- [30] Forbes K., Miltsakaki E., Prasad R., et al. D-LTAG system: Discourse parsing with a lexicalized tree-adjoining grammar[J]. Journal of Logic, Language and Information, 2003, 12(3): 261-279.
- [31] Wolf F., Gibson E. Representing discourse coherence: a corpus-based analysis[C]. In Proceedings of the 20th international conference on Computational Linguistics, 2004: 134-140.
- [32] Carlson L., Marcu D., and Okurowski M. E. Building a discourse-tagged corpus in the framework of rhetorical structure theory[M]. Springer Netherlands, 2003.
- [33] Carlson L., Okurowski M. E., and Marcu D. RST discourse Treebank[M]. Linguistic Data Consortium, University of Pennsylvania, 2002.
- [34] Hernault H., Bollegala D., and Ishizuka M. A Sequential Model for Discourse Segmentation[C]. In Computational Linguistics and Intelligent Text Processing, 2010: 315-326.
- [35] LeThanh H., Abeysinghe G., and Huyck C. Generating discourse structures for written texts[C]. In Proceedings of the 20th international conference on Computational Linguistics, 2004: 329.
- [36] Duverle D. A. and Prendinger H. A novel discourse parser based on support vector machine classification[C]. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009: 665-673.
- [37] 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.
- [38] Feng V. W. and Hirst G. Text-level discourse parsing with rich linguistic features[C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012: 60-68.
- [39] Joty S., Carenini G., and Ng R. A Novel Discriminative Framework for Sentence-

- Level Discourse Analysis[C]. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 904-915.
- [40] Joty S., Carenini G., Ng R., et al. Combining intra- and multisentential rhetorical parsing for document-level discourse analysis[C]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013:486–496.
- [41] Feng V.W. and Hirst G. A linear-time bottom-up discourse parser with constraints and post-editing[C]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014:511–521.
- [42] Ji Y.F. and Eisenstein J. Representation Learning for Text-level Discourse Parsing[C]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014:13-24.
- [43] Wellner B. and Pustejovsky J. Automatically Identifying the Arguments of Discourse Connectives[C]. In Proceedings of the EMNLP-CoNLL, 2007: 92-101.
- [44] Elwell R. and Baldridge J. Discourse connective argument identification with connective specific rankers[C]. In Proceedings of the 2008 IEEE International Conference on Semantic Computing, 2008:198-205.
- [45] Wellner B. Sequence models and ranking methods for discourse parsing[D]. Faculty of the Graduate School of Arts and Sciences Brandeis University Computer Science James Pustejovsky, Brandeis University, 2009.
- [46] Prasad R., Joshi A. K., and Webber B. L. Exploiting Scope for Shallow Discourse Parsing[C]. In Proceedings of the Seventh International Conference on Language Resources and their Evaluation, 2010:2076-2083.
- [47] 徐凡.英语篇章结构分析关键问题研究[D]. 苏州大学博士论文, 2013.
- [48] Dinesh N., Lee A., Miltsakaki E., et al. Attribution and the (non-)alignment of the Syntactic and Discourse Arguments of Connectives[C]. In Proceedings of ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky, 2005:29-36.
- [49] Pitler E., Raghupathy M., Mehta H., et al. Easily identifiable discourse relations[C]. In Proceedings of COLING, 2008:85-88.

- [50] Pitler E., Louis A., and Nenkova A. Automatic sense prediction for implicit discourse relations in text[C]. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2009: 683-691.
- [51] Lin Z.H., Kan M. Y., and Ng H. T. Recognizing implicit discourse relations in the Penn Discourse Treebank[C]. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009: 343-351.
- [52] Wang W. T., Su J., and Tan C. L. Kernel based discourse relation recognition with temporal ordering information[C]. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 710-719.
- [53] 徐凡, 朱巧明, 周国栋. 基于树核的隐式篇章关系识别[J]. 软件学报, 2013, 24(5): 1022-1035.
- [54] Fisher and Simmons. Spectral Semi-Supervised Discourse Relation Classification[C]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing(short paper), 2015:89-93.
- [55] Lin Z.H, Ng H.T and Kan M.Y. A PDTB-Styled End-to-End Discourse Parser[J]. Natural Language Engineering, 2014,20(2):151-184.
- [56] Hernault H., Bollegala D., and Ishizuka M. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension[C]. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010: 399-409.
- [57] 郑贵友. 汉语篇章分析的兴起与发展[J]. 汉语学习, 2005 (5): 40-48.
- [58] 吕叔湘. 汉语语法分析问题[M]. 北京: 商务印书馆, 1979.
- [59] 曹政. 句群初探[M]. 浙江: 浙江教育出版社, 1984.
- [60] 吴为章, 田小琳. 句群[M]. 上海: 上海教育出版社, 1984.
- [61] 乐明. 汉语财经评论的修辞结构标注及篇章研究[D]. 北京: 中国传媒大学博士学位论文, 2006.
- [62] 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.

- [63] 陈莉萍. 英汉语篇结构标注理论与实践[D]. 上海: 上海外国语大学, 2007.
- [64] Xue N.W. Annotating the Discourse Connectives in the Chinese Treebank[C]. In Proceedings of the ACL Workshop on Frontiers in Corpus An-notation, 2005:84-91.
- [65] Zhou Y.P. and Xue N.W. PDTB-style discourse annotation of Chinese text[C]. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012: 69-77.
- [66] Huang H.H. and Chen H.H. Chinese discourse relation recognition[C]. In Proceedings of the 5th International Joint Conference on Natural Language Processing, 2011:1442-1446.
- [67] Huang H.H. and Chen H.H. Contingency and comparison relation labeling and structure prediction in Chinese sentences[C]. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2012:261-269.
- [68] Huang H. H. and Chen H. H. An Annotation System for Development of Chinese Discourse Corpus[C]. In COLING (Demos), 2012: 223-230.
- [69] Zhou, L. J., Li B. Y., Wei Z. Y., et al. The CUHK Discourse TreeBank for Chinese: Annotating Explicit Discourse Connectives for the Chinese TreeBank[C]. In Proceedings of the International Conference on Language Resources and Evaluation, 2014:942-949
- [70] 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.
- [71] 邢福义. 汉语复句研究[M], 北京: 商务印书馆, 2001.
- [72] 姚双云. 复句关系标记的搭配研究及相关解释[D]. 武汉: 华中师范大学博士论文, 2006.
- [73] 周强. 汉语句法树库标注体系[J]. 中文信息学报, 2004, 18(4):1-8.
- [74] Xue N.W. and Yang Y.Q. Chinese sentence segmentation as comma classification[C]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (short paper), 2011: 631-635.
- [75] Yang Y.Q. and Xue N.W. Chinese comma disambiguation for discourse analysis[C]. In Proceedings of the 50th Annual Meeting of the Association for Computational

- Linguistics, 2012: 786-794.
- [76] 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.
- [77] 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.
- [78] 宋柔. 汉语篇章广义话题结构的流水模型[J]. 中国语文, 2013 (6): 483-494.
- [79] 胡胜高. 语言与外语教学多维研究[M]. 成都: 四川大学出版社, 2010.
- [80] 张军平. 英汉翻译中连接手段的使用差异及成因[J]. 西安外国语大学学报, 2008, 16(4):51-55.
- [81] Givon T. Topic continuity in discourse[M]. Amsterdam: John Benjamins.1983.
- [82] Sacks H. Schegloff E. A., and Jefferson G. A simplest systematic for the organization of turn-taking in conversation[J]. Language, 50, 1974:696-735.
- [83] Polanyi M. Personal knowledge: Towards a post-critical philosophy[M]. Psychology Press, 1998.
- [84] 王文格. 现代汉语小句的研究现状及存在的问题[J]. 汉语学习, 2010(1): 67-76.
- [85] 中华人民共和国国家标准. GB/T 15834—2011, 标点符号用法[S]. 中国标准出版社, 2012.
- [86] 周小佩, 洪宇, 车婷婷, 等. 一种无指导的隐式篇章关系推理方法研究[J]. 中文信息学报, 2013, 27(2):17-25.
- [87] 黄伯荣, 彦序东. 现代汉语(下册) [M]. 北京: 高等教育出版社, 2002.
- [88] Manning C.D. 等著, 苑春法等译. 统计自然语言处理基础[M]. 北京: 电子工业出版社, 2005.
- [89] 郭署纶. 汉语语料库的建设和使用[M]. 上海: 上海外语教育出版社, 2011.
- [90] 郑家恒, 张虎, 谭红叶, 等. 智能信息处理—汉语语料加工技术及应用[M]. 北京: 科学出版社, 2010.
- [91] Cohen J. A coefficient of agreement for nominal scales[J]. Educational and Psychological Measurement, 1960, 20 (1): 37-46.
- [92] 胡金柱, 吴锋文, 李琼, 等. 汉语复句关系词库的建设及其利用[J]. 语言科学, 2010, 9(2):133-142.

- [93] 胡金柱, 舒江波, 姚双云, 等. 面向中文信息处理的复句关系词提取算法研究[J]. 计算机工程与科学, 2009, 31(10):90-93.
- [94] 胡金柱, 陈江曼, 杨进才, 等. 基于规则的连用关系标记的自动标识研究[J]. 计算机科学, 2012, 39(7):190-194.
- [95] 洪鹿平. 汉语复句关系自动判断研究[D]. 南京: 南京师范大学. 2008.
- [96] 王东波, 陈小荷, 年洪东. 基于条件随机场的有标记联合结构自动识别[J]. 中文信息学报, 2008, 22(6):3-8.
- [97] 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.
- [98] Abney S., Flickenger S., Gdaniec C., et al. Procedure for quantitatively comparing the syntactic coverage of English grammars[C]. In Proceedings of the workshop on Speech and Natural Language, 1991: 306-311.

攻读博士学位期间完成的论文及科研工作

1. 论文

- [1] **Li Yancui**, Feng Wenhe, Sun Jing, Kong Fang, *Zhou Guodong. Building Chinese Discourse Corpus with Connective-driven Dependency Tree Structure[C]. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP 2014), 2014:2105–2114. (CCF 推荐国际学术会议目录 B 类会议 EI Indexed: 20151400707960)
- [2] **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(2):307-314.
- [3] **李艳翠**, 谷晶晶, *周国栋. 添加冒号和分号分类特征的逗号分类[J]. 中文信息学报, 2014, 28(5):284-291.
- [4] **李艳翠**, 冯文贺, *周国栋, 朱坤华. 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.
- [5] **Li Yancui**, Feng Hongyu, Feng Wenhe. Chinese discourse segmentation based on punctuation marks [J]. International Journal of Signal Processing, Image Processing and Pattern Recognition, 2015, 8(3):177-186. (EI Indexed: 20151400720278)
- [6] **李艳翠**, 孙静, 冯文贺, *周国栋. 基于连接依存树的汉语篇章结构分析平台[J]. 全国第十四届计算语言学会议(CCL 2015), 已录用.
- [7] **李艳翠**, 孙静, *周国栋, 冯文贺. 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.
- [8] **Li Yancui**, Feng Wenhe, *Zhou Guodong. Elementary discourse unit in Chinese discourse structure analysis[C]. In *Chinese Lexical Semantics*, 2013, pages 186-198, Wuhan, China, Springer Berlin Heidelberg. (EI Indexed: 20130916058300)
- [9] **李艳翠**, 朱坤华, *周国栋. 英语语篇结构分析研究综述[J]. 计算机应用研究, 2012, 29(6):2018-2023.
- [10] **李艳翠**, 林莉媛, *周国栋. 基于有监督学习方法的多文档文本情感摘要[J]. 中文信息学报, 2014, 28(6):426-432.

- [11] 孙静, **李艳翠**, *周国栋, 冯文贺. 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.
- [12] 冯文贺, **李艳翠**, *周国栋. 汉英篇章结构平行语料库的对齐标注评估[J]. 中文信息学报, 已录用.
- [13] 冯洪玉, **李艳翠**, *周国栋. 基于汉英平行语料库的英文显式篇章关系识别[J]. 中文信息学报, 已录用.

2. 软件著作权登记

- [1] 周国栋, 李艳翠. 汉语篇章结构标注平台. 登记号: 2014SR076349.
- [2] 冯文贺, 李艳翠. 汉英篇章结构平行语料库对齐标注软件. 登记号:2014SR171182.

3. 主持科研项目

- [1] 汉英篇章衔接对齐资源构建与分析研究, 国家自然科学基金青年科学基金项目(61502149), 2016.01-2018.12

4. 骨干参与项目

- [1] 汉语篇章结构分析的资源建设与计算模型研究. 国家自然科学基金面上项目(61273320), 2013.01-2016.12
- [2] 面向中文语义计算的概念分类体系及指代消歧方法研究. 国家高技术研究发展计划(863 计划)子课题(2012AA011102), 2012.01-2014.12

附录

附表1. 连接词所对应的篇章关系及频次

序号	连接词	频次	关系
1	并	563	顺承关系(16)递进关系(1)并列关系(546)
2	因此	401	因果关系(401)
3	并且	266	顺承关系(1)并列关系(265)
4	其中	201	总分关系(201)
5	也	164	顺承关系(2)并列关系(162)
6	例如	147	并列关系(1)例证关系(146)
7	而	117	递进关系(6)顺承关系(1)转折关系(6)对比关系(33)并列关系(70)因果关系(1)
8	以	114	解说关系(1)目的关系(113)
9	还	83	顺承关系(12)并列关系(70)选择关系(1)
10	但	79	转折关系(76)对比关系(3)
11	来	68	目的关系(68)
12	然后	57	顺承关系(57)
13	使	57	因果关系(40)目的关系(17)
14	因为	53	因果关系(53)
15	同时	48	并列关系(48)
16	为	47	目的关系(47)
17	后	45	顺承关系(45)
18	由于	38	因果关系(38)
19	但是	33	转折关系(33)
20	如	31	条件关系(1)例证关系(18)假设关系(12)
21	如果	30	假设关系(30)
22	而且	27	顺承关系(1)递进关系(1)并列关系(25)
23	又	27	顺承关系(10)并列关系(17)
24	为了	23	目的关系(23)
25	然而	23	转折关系(18)对比关系(4)并列关系(1)
26	此外	16	并列关系(16)
27	後	16	顺承关系(16)
28	从而	16	因果关系(16)
29	虽然...但	16	转折关系(16)
30	亦	12	顺承关系(1)并列关系(10)选择关系(1)
31	另外	11	并列关系(11)

序号	连接词	频次	关系
32	既…也	10	并列关系(10)
33	尽管…但	10	让步关系(10)
34	为此	9	因果关系(8)目的关系(1)
35	更	9	递进关系(8)并列关系(1)
36	虽…但	9	转折关系(9)
37	由於	9	因果关系(9)
38	以及	8	并列关系(8)
39	若	8	假设关系(8)
40	同时也	8	并列关系(8)
41	以便	8	目的关系(8)
42	不过	7	转折关系(6)对比关系(1)
43	同时…也	7	并列关系(7)
44	即	7	解说关系(7)
45	与此同时	6	并列关系(6)
46	甚至	6	递进关系(6)
47	不管	6	条件关系(6)
48	且	6	并列关系(6)
49	对此	6	因果关系(3)背景关系(3)
50	而…也	6	并列关系(6)
51	尤其	6	递进关系(6)
52	同时…还	6	并列关系(6)
53	不仅…而且	6	递进关系(5)并列关系(1)
54	此外…还	6	并列关系(6)
55	除…外…还	6	并列关系(6)
56	如果…那么	5	假设关系(5)
57	或	5	并列关系(1)选择关系(4)
58	让	5	目的关系(5)
59	加上	5	并列关系(5)
60	只要	5	条件关系(5)
			转折关系(1)条件关系(1)假设关系(2)推
61	否则	5	断关系(1)
62	因而	4	因果关系(4)
63	虽然…但是	4	转折关系(4)
64	自…以来	4	顺承关系(3)背景关系(1)
65	不是…而是	4	并列关系(4)
66	除…外	4	并列关系(4)
67	不仅…也	4	并列关系(4)

序号	连接词	频次	关系
68	却	4	转折关系(3)并列关系(1)
69	接着	4	顺承关系(4)
70	如今	4	对比关系(4)
71	既…又	4	并列关系(4)
72	此后	4	顺承关系(4)
73	此外…也	4	并列关系(4)
74	以使	4	目的关系(4)
75	由此	4	因果关系(4)
76	如果…就	4	假设关系(4)
77	但…却	4	转折关系(4)
78	随后	4	顺承关系(4)
79	以期	4	目的关系(4)
80	所以	4	因果关系(4)
81	只有…才	4	条件关系(4)
82	与此同时…也	3	并列关系(3)
83	之所以	3	因果关系(3)
84	使得	3	因果关系(3)
85	和	3	并列关系(3)
86	一…二…三…四	3	并列关系(3)
87	以此	3	目的关系(3)
88	假如	3	假设关系(3)
89	加之	3	并列关系(3)
90	才	3	条件关系(2)并列关系(1)
91	鉴於	3	因果关系(3)
92	不仅…而且还	3	递进关系(3)
93	之所以…是因为	3	因果关系(3)
94	不仅…还	3	并列关系(3)
95	除…外…并	3	并列关系(3)
96	时	3	条件关系(2)并列关系(1)
97	由於…因此	3	因果关系(3)
98	然而…却	3	转折关系(3)
99	而…又	3	对比关系(1)并列关系(2)
100	于是	3	因果关系(2)顺承关系(1)
101	除了…外…还	2	并列关系(2)
102	另外…还	2	并列关系(2)
103	最后	2	顺承关系(2)
104	至此	2	解说关系(1)总分关系(1)

序号	连接词	频次	关系
105	与此同时…还	2	并列关系(2)
106	由于…因此	2	因果关系(2)
107	在…后	2	顺承关系(2)
108	可是	2	转折关系(2)
109	而…却	2	转折关系(1)对比关系(1)
110	何况	2	递进关系(2)
111	同样	2	并列关系(2)
112	正因为	2	因果关系(2)
113	因为…因为…由於	2	因果关系(2)
114	意味着	2	评价关系(2)
115	过去…现在	2	对比关系(2)
116	因	2	因果关系(2)
117	尽管…但是	2	让步关系(2)
118	特别是	2	递进关系(1)总分关系(1)
119	而且…也	2	并列关系(2)
120	同时还	2	并列关系(2)
121	除了…外	2	并列关系(2)
122	終於	2	顺承关系(2)
123	以来	2	条件关系(1)顺承关系(1)
124	不论	2	条件关系(2)
125	进一步	2	递进关系(2)
126	一旦	2	条件关系(1)假设关系(1)
127	再	2	顺承关系(1)并列关系(1)
128	而…更	2	递进关系(2)
129	不仅…而且…也	2	并列关系(2)
130	由於…故	2	因果关系(2)
131	以后	2	顺承关系(2)
132	在…时	2	条件关系(1)顺承关系(1)
133	先…然后	1	顺承关系(1)
134	与其…不如	1	选择关系(1)
135	在…的情况下	1	条件关系(1)
136	当…後	1	条件关系(1)
137	不过…却	1	转折关系(1)
138	旨在	1	目的关系(1)
139	或者…再说…再	1	并列关系(1)
140	不是…却是	1	并列关系(1)
141	一旦…就	1	条件关系(1)

序号	连接词	频次	关系
142	在…之前	1	条件关系(1)
143	尽管	1	让步关系(1)
144	为了…为了	1	目的关系(1)
145	也…而	1	并列关系(1)
146	比如	1	例证关系(1)
147	正如…一样	1	例证关系(1)
148	进而	1	递进关系(1)
149	从此以后	1	顺承关系(1)
150	一…二…三	1	并列关系(1)
151	当时	1	并列关系(1)
152	以免	1	目的关系(1)
153	但是…却	1	转折关系(1)
154	除了…之外	1	并列关系(1)
155	最终	1	顺承关系(1)
156	虽然…但也	1	转折关系(1)
157	随后	1	顺承关系(1)
158	由于…所以	1	因果关系(1)
	首先…其次…第三…		
159	最后	1	并列关系(1)
160	将	1	并列关系(1)
	不仅是…不仅是…也		
161	不仅是…最重要的是	1	递进关系(1)
162	正是如此	1	因果关系(1)
163	即使	1	让步关系(1)
164	加上…也	1	并列关系(1)
165	使之	1	目的关系(1)
166	虽…却	1	转折关系(1)
167	一是…二是…三是	1	并列关系(1)
168	紧接	1	顺承关系(1)
169	最重要的	1	递进关系(1)
170	於是…随即	1	顺承关系(1)
171	不再…只是	1	并列关系(1)
172	如…便	1	条件关系(1)
173	如果…便	1	假设关系(1)
174	虽然…但却	1	转折关系(1)
175	其中包括	1	总分关系(1)
176	首先	1	顺承关系(1)

序号	连接词	频次	关系
177	只见…不见	1	并列关系(1)
178	主要…其次	1	并列关系(1)
179	即使…还	1	让步关系(1)
	一…二…三…四…		
180	五…六	1	并列关系(1)
181	也…还	1	并列关系(1)
182	首先…再来	1	顺承关系(1)
183	另加	1	并列关系(1)
184	再说…也	1	并列关系(1)
185	相较之下	1	对比关系(1)
186	例	1	例证关系(1)
187	当时…现在	1	对比关系(1)
188	过去…如今	1	对比关系(1)
189	既不…也不…而是	1	并列关系(1)
190	等於是	1	评价关系(1)
191	然而…又	1	对比关系(1)
192	正因	1	因果关系(1)
193	只要…就	1	条件关系(1)
194	然而…也	1	转折关系(1)
195	自此	1	顺承关系(1)
196	因为…因此	1	因果关系(1)
197	同时亦	1	并列关系(1)
198	除了…还	1	并列关系(1)
	不是…不是…而是…		
199	是	1	并列关系(1)
200	也可以说	1	推断关系(1)
201	突然	1	转折关系(1)
202	只…就	1	条件关系(1)
203	此时	1	并列关系(1)
204	也…此外	1	并列关系(1)
205	后…还	1	顺承关系(1)
206	紧接着…又	1	顺承关系(1)
207	既然	1	转折关系(1)
	(甲)…(乙)…(丙)…		
208	及(丁)	1	并列关系(1)
209	或者	1	选择关系(1)
210	也就是说	1	解说关系(1)

序号	连接词	频次	关系
211	不仅…更	1	递进关系(1)
212	不应该…而应该	1	并列关系(1)
213	故	1	因果关系(1)
214	及	1	并列关系(1)
215	诸如	1	例证关系(1)
216	在这同时…还	1	并列关系(1)
217	一是	1	总分关系(1)
218	使…以	1	目的关系(1)
219	亦…或	1	选择关系(1)
220	虽是…但…也是	1	并列关系(1)
221	也…并	1	并列关系(1)
222	如果…的话	1	假设关系(1)
223	还…同时…还	1	并列关系(1)
224	无论	1	条件关系(1)
225	幸而	1	转折关系(1)
226	之后	1	顺承关系(1)
227	尤其是	1	递进关系(1)
228	也是	1	并列关系(1)
229	(一) …及 (二)	1	并列关系(1)
230	既…又…还	1	并列关系(1)
231	唯有…才	1	条件关系(1)
232	今后	1	顺承关系(1)
233	看来	1	推断关系(1)
234	继而	1	顺承关系(1)
235	不久	1	顺承关系(1)
236	也…而且	1	并列关系(1)
237	除了…外…同时	1	并列关系(1)
	一是…二是…三是…		
238	四是	1	并列关系(1)
239	一方面…一方面也	1	并列关系(1)
240	更进一步	1	递进关系(1)
241	虽然…但…还	1	转折关系(1)
242	过去…现今	1	对比关系(1)
243	一到	1	条件关系(1)
244	一是…二是	1	并列关系(1)
245	一旦…就…就	1	条件关系(1)
246	足见	1	推断关系(1)

序号	连接词	频次	关系
247	於是	1	顺承关系(1)
248	由于…因而	1	因果关系(1)
249	如果…则	1	假设关系(1)
250	正当	1	并列关系(1)
251	在…後	1	顺承关系(1)
252	还有	1	并列关系(1)
253	或…或	1	选择关系(1)
254	还…也	1	并列关系(1)
255	另一方面…还	1	并列关系(1)
256	可以说	1	解说关系(1)
	一是…二是…三是…		
	四是…五是…五是…		
257	六是…七是	1	并列关系(1)
258	从此	1	因果关系(1)
259	紧接着	1	顺承关系(1)
260	譬如	1	例证关系(1)
261	并不…而更	1	递进关系(1)
262	据此	1	因果关系(1)
263	虽然…却又	1	转折关系(1)
264	那么	1	顺承关系(1)
265	更加	1	递进关系(1)
266	又…也	1	并列关系(1)
267	不过是	1	转折关系(1)
268	只是	1	转折关系(1)
269	不但…也	1	并列关系(1)
270	由於…故此	1	因果关系(1)
271	其後	1	顺承关系(1)
272	特别是…又	1	递进关系(1)
273	而…加之	1	并列关系(1)
274	不能…只能	1	并列关系(1)
275	换句话说	1	并列关系(1)
276	自…后	1	顺承关系(1)
277	虽然…仍	1	让步关系(1)
278	此后…又	1	顺承关系(1)

附表2. 篇章关系所对应的连接词及频次

篇章关系	频次	连接词及频次
并列关系	3507	<p>(2080)与此同时…也(3)与此同时(6)不仅…而且(1)不是…却是(1)例如(1)也…而(1)另外…还(2)以及(8)一…二…三(1)当时(1)且(6)不仅…而且…也(2)除了…之外(1)而(70)或者…再说…再(1)将(1)与此同时…还(2)而…也(6)此外(16)一是…二是…三是(1)时(1)除…外…还(6)不再…只是(1)此外…还(6)只见…不见(1)主要…其次(1)也…还(1)同时也(8)另加(1)再说…也(1)还(70)不是…而是(4)而且(25)更(1)此时(1)除…外(4)同时亦(1)除…外…并(3)不是…不是…而是…是(1)除了…外(2)却(1)同时…还(6)也…此外(1)加上(5)加之(3)并且(265)(甲)…(乙)…(丙)…及(丁)(1)亦(10)既…也(10)才(1)不应该…而应该(1)一是…二是…三是…四是(1)除了…还(1)在这同时…还(1)而且…也(2)加上…也(1)并(546)虽是…但…也是(1)同时还(2)既…又(4)也…并(1)还…同时…还(1)也(162)另一方面…还(1)此外…也(4)又…也(1)也是(1)既…又…还(1)除了…外…还(2)不仅…还(3)和(3)又(17)也…而且(1)除了…外…同时(1)同时(48)一方面…一方面也(1)或(1)同时…也(7)及(1)一是…二是(1)不能…只能(1)再(1)正当(1)另外(11)还有(1)不仅…也(4)一是…二是…三是…四是…五是…五是…六是…七是(1)一…二…三…四(3)然而(1)而…又(2)首先…其次…第三…最后(1)既不…也不…而是(1)(一)…及(二)(1)不但…也(1)同样(2)而…加之(1)一…二…三…四…五…六(1)换句话说(1)还…也(1)</p>
解说关系	907	(896)以(1)即(7)可以说(1)也就是说(1)至此(1)
因果关系	686	<p>(69)由於(9)使得(3)之所以(3)为此(8)由于(38)由于…因而(1)故(1)因(2)鑑於(3)正是如此(1)由于…因此(2)正因为(2)由於…因此(3)从此(1)正因(1)据此(1)对此(3)之所以…是因为(3)由於…故(2)因而(4)而(1)由于…所以(1)因为…因此(1)所以(4)由此(4)使(40)因为…因为…由於(2)因此(401)因为(53)于是(2)从而(16)由於…故此(1)</p>
顺承关系	515	<p>(304)先…然后(1)并且(1)終於(2)紧接(1)今后(1)亦(1)於是(1)於是…随即(1)在…後(1)最后(2)再(1)其後(1)从此以后(1)接着(4)首先…再来(1)并(16)隨後(1)随后(4)此后(4)紧接着…又(1)以来(1)也(2)紧接着(1)还(12)而且(1)然后(57)之后(1)最终(1)在…时(1)那么(1)而(1)自此(1)后(45)以后(2)继而(1)在…后(2)于是(1)不久(1)又(10)首先(1)自…后(1)自…以来(3)後(16)此后…又(1)后…还(1)</p>

篇章关系	频次	连接词及频次
目的关系	333	(35)以(113)为(47)使之(1)使(17)以免(1)为了(23)以便(8)为了…为了(1)使…以(1)为此(1)旨在(1)来(68)以使(4)让(5)以此(3)以期(4)
例证关系	253	(84)诸如(1)比如(1)正如…一样(1)譬如(1)如(18)例如(146)例(1)
总分关系	235	(30)其中包括(1)特别是(1)其中(201)至此(1)一是(1)
评价关系	221	(218)等於是(1)意味着(2)
转折关系	196	(1)但是(33)虽然…但是(4)虽…却(1)而…却(1)既然(1)幸而(1)不过(6)虽然…但却(1)但…却(4)然而(18)然而…却(3)虽…但(9)虽然…但(16)虽然…却又(1)可是(2)但是…却(1)否则(1)虽然…但也(1)而(6)不过是(1)只是(1)突然(1)不过…却(1)却(3)然而…也(1)但(76)虽然…但…还(1)
背景关系	134	(130)自…以来(1)对此(3)
条件关系	71	(34)当…後(1)一到(1)不论(2)一旦…就…就(1)一旦…就(1)一旦(1)在…之前(1)无论(1)不管(6)在…的情况下(1)如(1)以来(1)只要(5)否则(1)才(2)在…时(1)唯有…才(1)只要…就(1)如…便(1)时(2)只有…才(4)只…就(1)
假设关系	69	(1)如果…那么(5)如果…的话(1)假如(3)否则(2)如果…便(1)如果…就(4)如果(30)若(8)一旦(1)如(12)如果…则(1)
对比关系	60	(6)然而(4)过去…如今(1)不过(1)而…又(1)而…却(1)然而…又(1)过去…现今(1)当时…现在(1)如今(4)但(3)而(33)过去…现在(2)相较之下(1)
递进关系	59	(7)甚至(6)更加(1)进一步(2)特别是(1)进而(1)而…更(2)不仅…而且还(3)并(1)最重要的(1)并不…而更(1)而且(1)尤其是(1)更(8)而(6)不仅…而且(5)不仅…更(1)尤其(6)不仅是…不仅是…也不仅是…最重要的是(1)特别是…又(1)何况(2)更进一步(1)
推断关系	38	(34)看来(1)也可以说(1)足见(1)否则(1)
让步关系	16	尽管(1)尽管…但(10)即使(1)即使…还(1)尽管…但是(2)虽然…仍(1)
选择关系	10	或(4)与其…不如(1)还(1)或者(1)亦…或(1)亦(1)或…或(1)

致 谢

年年岁岁花相似，岁岁年年人不同，6年的博士生涯终将落幕。回首多年的学习生活，百感交集，在此向一直帮助我的人们表示衷心的感谢。

桃李不言，下自成蹊，感谢我的导师周国栋教授！学术上，周老师对我悉心指导，小论文的发表，大论文的开题、研究、写作，周老师始终为我把握方向；生活上，周老师对我宽容、理解、帮助和支持，虽然有压力，但您仍然让我体会到了生活的丰富多彩；思想上，周老师严肃对事、宽厚待人的作风，周老师的师德和专业的造诣值得我终生学习。在此谨向周老师致以诚挚的谢意和崇高的敬意，希望您身体健康、万事如意。高山不移，碧水常流，我师恩泽，在心永留！

学贵得师，亦贵得友，感谢自然语言处理实验室的朱巧明老师、张民老师、钱龙华老师、孔芳老师、李寿山老师、贡正仙老师、王红玲老师、李培峰老师、洪宇老师、李军辉老师、李正华老师、段湘煜老师，同您们的交流总能让我我受益良多。

海内存知己，天涯若比邻，感谢实验室朝夕相处的同学们。特别感谢徐凡、孙静、博伟、中卿、雪峰、许兰、黄鑫、晓敏、朱珠、谷晶晶、林莉媛、方艳、戴敏等等，以及这些年我看着你们入学又看着你们毕业的师弟师妹们，正是你们的陪伴，给我留下了许多充实而美好的回忆，让我的博士生生活多姿多彩，充满着集体的温暖。

感谢学院领导和同事给我工作学习上提供了便利条件。特别感谢冯文贺博士，感谢你以语言学的视角对我研究工作及论文提出的宝贵意见和建议。

感谢我的前任室友欧阳芬老师，现任室友钟姍，与你们在一起生活的时光是美好的，晚上的卧谈总是轻松而难忘的幸福时刻。

当然更要感谢我的家人，特别感谢婆婆这几年为家里的辛苦付出，没有您的帮助，我将无法照顾工作、学习和家庭。感谢爸爸的理解和支持，看着您日渐增多的白发，我知道作为女儿我做的远远不够。感谢老公对我的支持和宽容。感谢我5岁的女儿，你是上天赐予我最好的礼物，妈妈以后会好好陪你，希望我的宝贝能永远开心快乐。

感谢公开学习资源的专家学者以及参考文献的作者和编者，本文的完成离不开你们的无私奉献和辛勤工作。

最后，再次感谢曾经关心和帮助我的人们。

衷心感谢在百忙之中参加评审的各位专家学者，烦请对论文的偏颇与疏漏之处不吝批评指正。