

# 英汉篇章结构分析研究综述\*

蒋峰<sup>1,2</sup>, 范亚鑫<sup>1,2</sup>, 褚晓敏<sup>1,2</sup>, 李培峰<sup>1,2</sup>, 朱巧明<sup>1,2</sup>

<sup>1</sup>(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

<sup>2</sup>(江苏省大数据智能工程实验室, 江苏 苏州 215006)

通信作者: 朱巧明, E-mail: [qmzhu@suda.edu.cn](mailto:qmzhu@suda.edu.cn)



**摘要:** 篇章结构分析旨在理解文章的整体结构及其各部分之间的语义联系. 作为自然语言处理的研究热点, 近年来篇章结构分析研究发展迅速. 首先总结英语和汉语中篇章结构分析理论, 然后介绍相关篇章语料库及其计算模型的研究. 在此基础上, 梳理了当前英语、汉语中篇章结构分析的相关工作脉络, 构建了篇章结构分析研究框架, 归纳总结出当前研究的趋势和热点. 然后, 简要介绍篇章结构在下游任务中的应用. 最后, 指出当前汉语篇章结构分析存在的问题与挑战, 为今后的研究提供指导和帮助.

**关键词:** 自然语言处理; 篇章分析; 篇章结构; 篇章关系; 篇章语料库

**中图法分类号:** TP18

中文引用格式: 蒋峰, 范亚鑫, 褚晓敏, 李培峰, 朱巧明. 英汉篇章结构分析研究综述. 软件学报, 2023, 34(9): 4167-4194. <http://www.jos.org.cn/1000-9825/6650.htm>

英文引用格式: Jiang F, Fan YX, Chu XM, Li PF, Zhu QM. Survey on English and Chinese Discourse Structure Analysis. Ruan Jian Xue Bao/Journal of Software, 2023, 34(9): 4167-4194 (in Chinese). <http://www.jos.org.cn/1000-9825/6650.htm>

## Survey on English and Chinese Discourse Structure Analysis

JIANG Feng<sup>1,2</sup>, FAN Ya-Xin<sup>1,2</sup>, CHU Xiao-Min<sup>1,2</sup>, LI Pei-Feng<sup>1,2</sup>, ZHU Qiao-Ming<sup>1,2</sup>

<sup>1</sup>(School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

<sup>2</sup>(Engineering Laboratory of Big Data Intelligence of Jiangsu Province, Suzhou 215006, China)

**Abstract:** Discourse structure analysis aims to understand the overall structure of an article and the semantic relationships between its various parts. As a research hotspot of natural language processing, it has developed rapidly in recent years. This study first summarizes the mainstream discourse structure analysis theories in English and Chinese and then introduces the research on the popular English and Chinese discourse corpora as well as their calculation models. On this basis, this study surveys the current work context of discourse structure analysis in Chinese and English and constructs its research framework. Moreover, the current research trends and focuses are summarized, and the application of discourse structure in downstream tasks is introduced briefly. Finally, this study points out the issues and challenges in the current Chinese discourse structure analysis to provide guidance and help for future research.

**Key words:** natural language processing (NLP); discourse analysis; discourse structure; discourse relation; discourse corpus

## 1 引言

在自然语言处理领域, 随着研究对象逐渐从细颗粒度的字、词和句等向粗颗粒度的段落和篇章等转移, 篇章结构分析逐渐成为其中的一个研究热点. 篇章不仅是文本序列本身, 而且是由连续的句子或段落构成的语言整体. 因此, 无论在形式还是意义上, 篇章都不是孤立存在, 而是由各自承担一定角色的篇章单元组织成相互关联的篇章

\* 基金项目: 国家自然科学基金 (61836007, 61772354, 61773276); 江苏高校优势学科建设工程项目  
收稿时间: 2021-08-22; 修改时间: 2021-12-12; 采用时间: 2022-01-26; jos 在线出版时间: 2022-05-24  
CNKI 网络首发时间: 2023-02-08

结构,并通过篇章关系关联在一起,从而表达特定的篇章语义和意图.篇章结构分析的目的是在理解文本语义的基础上,分析篇章内在、固有的结构.作为篇章结构分析的研究对象,篇章单元根据颗粒度不同可以是子句、句子、段落或篇章.由于段落内部和段落之间的篇章结构特点不同,一般的,将段落内部的篇章结构称为微观篇章结构,而段落之间的篇章结构称为宏观篇章结构.

撰写一篇高质量的文章应该从遣词造句和谋篇布局两个部分考虑,如图 1 所示.遣词造句是指在字词级别更好地表示文本语义.而谋篇布局是从更高的层次整合设计文章的篇章结构,从而完成从文本语义到作者思想的映射表达.一般而言,篇章结构分析理论将一篇文章的篇章结构表示为树<sup>[1-3]</sup>或者图<sup>[4,5]</sup>的形式,其目的是梳理文本序列间存在的语义联系,并从整体上把握其含义.研究篇章结构分析有助于更好地理解篇章的整体框架及其各部分作用,为许多自然语言处理下游应用,如情感分析<sup>[6-9]</sup>、自动摘要<sup>[10-17]</sup>和机器翻译<sup>[18-20]</sup>等提供帮助.

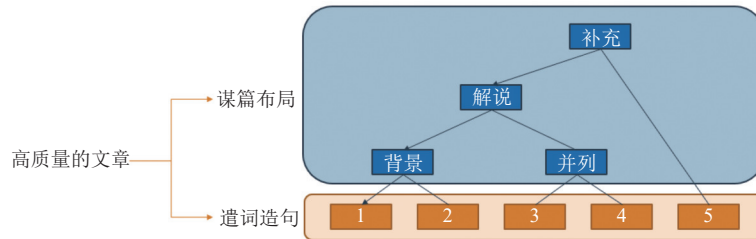


图 1 高质量文章的组成部分

最近,随着一系列预训练模型<sup>[21-25]</sup>在句子级任务中取得众多突破后,自然语言处理的研究重点逐渐向高于句子级别的任务转移.例如,一些研究尝试将实体<sup>[26]</sup>、知识图谱<sup>[27]</sup>、句法<sup>[28]</sup>等信息融入到预训练模型中.因此,在英语中,研究句子及以上层次的文章结构和关系的篇章结构分析领域发展迅速,尤其是在具有代表性的标注层次化篇章关系(hierarchical discourse relation)的修辞结构理论篇章树库(rhetorical structure theory discourse TreeBank, RST-DT)<sup>[29]</sup>和标注扁平化篇章关系(flat discourse relation)的宾州篇章树库(Penn discourse TreeBank, PDTB)<sup>[30]</sup>的研究日趋成熟.在汉语篇章语料库上的研究也逐渐取得进展,例如标注层次化篇章关系的语料库 SUDA-CDTB (SUDA Chinese discourse TreeBank)<sup>[31]</sup>和 MCDTB (macro-level Chinese discourse TreeBank)<sup>[31,32]</sup>和标注扁平化篇章关系的语料库 HIT-CDTB (HIT Chinese discourse TreeBank)<sup>[33]</sup>和 LDC-CDTB (LDC Chinese discourse TreeBank)<sup>[34]</sup>等.在其他语言中为篇章结构分析而构建的语料库也逐渐兴起,如德语<sup>[35]</sup>、荷兰语<sup>[36]</sup>、巴斯克语<sup>[37]</sup>、西班牙语<sup>[38]</sup>等.本文选取研究较为成熟和丰富的英语及汉语的篇章结构分析相关工作进行重点介绍.

目前,针对英语篇章结构分析研究的相关工作已有一些综述.Hou 等人<sup>[39]</sup>介绍了 RST-DT 语料库的理论、研究方法和应用.严为绒等人<sup>[40]</sup>针对篇章关系识别任务在 RST-DT, PDTB 和 HIC-CDTB 语料库上的工作进行了综述.Morey 等人<sup>[41]</sup>则针对现有的在 RST-DT 语料库上的计算模型进行了统一的评测,并提出了更加严格的新型篇章结构树评估标准.然而这些综述只专注于某一种篇章结构分析理论或某一篇章结构分析任务的相关工作.Song 等人<sup>[42]</sup>主要介绍了基于 RST-DT 语料库的篇章分析,基于 PDTB 语料库的篇章分析和连贯性评价 3 个方面的相关工作.Li 等人<sup>[43]</sup>主要介绍了基于 RST-DT 语料库的篇章分析和基于 PDTB 语料库的篇章分析相关工作,并介绍了关于多方对话的篇章分析研究工作.但是,以上工作都较少关注最近 3 年以来基于预训练模型和无监督半监督方法在篇章分析上的研究进展.

而在汉语中的研究工作一般聚焦于某些模型或者方法,还尚未对汉语篇章结构分析相关工作进行针对性、系统性的总结.2013 年, Xu 等人<sup>[44]</sup>综合描述汉语和英语的篇章分析技术的研究进展.当时汉语篇章分析技术研究主要集中在篇章结构分析理论和语料库建设部分,汉语篇章结构分析方法和应用还在探索之中.Kang 等人<sup>[45]</sup>只对于汉语篇章标注的篇章表示形式进行了综述,并未对篇章结构分析方法和应用进行介绍.孔芳等人<sup>[46]</sup>则针对汉语篇章理解研究,从篇章的连贯性和衔接性的角度对于篇章修辞结构和篇章话题结构进行了探讨,该文侧重于篇章结构理论方面工作的描述,对于具体的篇章结构解析方法和应用没有进行深入的归纳和梳理.

因此,有必要梳理英语篇章结构分析的研究脉络,特别是需要关注最近3年来在流行的层次化篇章关系语料库 RST-DT 和扁平化篇章关系语料库 PDTB 上的研究工作。同时,需要系统的总结当前汉语篇章结构分析的研究成果,尤其是近期在汉语篇章结构分析方法以及应用等方面取得的进展。通过对比英语和汉语篇章结构分析研究工作,指出汉语篇章结构分析中存在的问题与挑战,为下一步工作提供指导和帮助。为此,本文基于之前的综述研究,特别关注人工智能和自然语言处理国际会议(AAAI, IJCAI, ACL, EMNLP, COLING, NAACL, NLPCC等)以及国内重要期刊(软件学报,计算机学报,中文信息学报等)最近3-5年发表的篇章结构分析相关研究工作,对于英汉篇章结构分析研究进行综述介绍。

本文第2节对英汉篇章结构分析理论的相关工作进行介绍和梳理,并指出篇章结构表示的3种主要形式(浅层、树型和图型)。第3节介绍英语篇章结构分析方法,重点梳理层次化篇章关系语料库 RST-DT 和扁平化篇章关系语料库 PDTB 上的相关工作,并总结篇章结构分析的研究框架。第4节介绍汉语篇章结构分析方法,重点梳理层次化篇章关系语料库(SUDA-CDTB 和 MCDTB)和扁平化篇章关系语料库(HIT-CDTB 和 LDC-CDTB)上的相关工作。第5节简要介绍篇章结构分析在下游任务的应用,梳理其在文本理解和文本生成两个方向的进展情况。第6节对比英汉篇章结构分析研究的异同。第7节介绍篇章结构分析相关研究趋势,并主要指出汉语宏观篇章分析目前存在的问题和挑战以及探索接下来的研究方向。第8节对于全文进行总结和展望。

## 2 篇章结构分析理论

根据篇章结构分析理论作用范围及其表现形式,英语中主流的篇章结构分析理论可以分为单文档结构理论和多文档结构理论两大类,如图2所示。在研究单文档的篇章结构分析理论中,早期 Hobbs 模型<sup>[47]</sup>指出篇章结构的两个关键组成部分是篇章单元以及两个篇章单元间存在的关系,并且定义了12种篇章关系。后来随着理论体系的完善和发展,根据篇章结构的不同表示形式,篇章结构分析理论可分为树型结构表示的修辞结构理论(rhetorical structure theory, RST)<sup>[1]</sup>,浅层结构表示的篇章词汇化树衔接语法理论(discourse parsing with a lexicalized tree-adjoining grammar, D-LTAG)<sup>[48]</sup>和图型结构表示的分段式篇章表示理论(segment discourse representation theory, SDRT)<sup>[49]</sup>等。而涉及多文档篇章结构理论以跨文档结构理论(cross-document structure theory, CST)<sup>[4]</sup>为代表。在汉语中,早期以中国传统语言学篇章结构分析理论为主,如句群理论<sup>[50]</sup>和复句理论<sup>[51]</sup>等,后来研究逐渐与国际理论同步,如汉语框架语义网理论(Chinese FrameNet, CFN)<sup>[52]</sup>,基于连接依存树的汉语篇章结构理论(Chinese discourse structure theory based on connective-driven dependency tree)<sup>[2]</sup>,中文篇章级句间语义关系体系<sup>[33]</sup>,汉语宏观篇章结构表示理论<sup>[33,54]</sup>等。

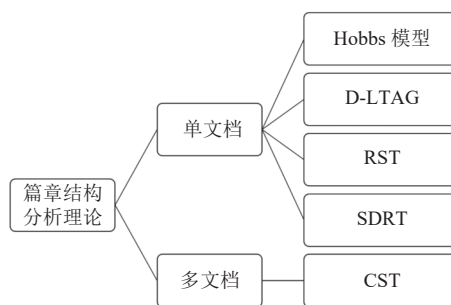


图2 英语篇章结构分析理论

### 2.1 英语篇章分析理论

- Hobbs 模型提出篇章单元以及两个篇章单元之间存在的连接关系是篇章结构的基本组成部分。其中,篇章单元可以是子句、句子和句群,甚至是篇章本身,而连接关系是指篇章单元之间的语义关联性。该理论定义了12类篇章关系,包括详述、并列、结果、背景和时机等。

- 修辞结构理论是一种基于树型结构的理论,早期应用于计算机的文本自动生成任务,是目前主流的篇章结

构和功能描述研究的理论之一。修辞结构理论与 Hobbs 模型具有很大的相似性, 同样认为篇章结构应该由篇章单元和篇章关系共同组成。但是, 修辞结构理论认为篇章结构具有层次性, 因此通过相邻的两个篇章单元及其间的篇章关系可以构成更大的篇章单元, 从而与外界发生语义联系, 最终形成层次化的树型结构, 如图 3 所示。该理论定义了 4 大类共 25 小类修辞关系, 每个关系可连接相邻的两个或多个篇章单元。并且, 该理论额外标识了篇章单元间的主次关系。如果修辞关系连接的篇章单元间存在主次, 那么包含重要信息的篇章单元称作“核 (nucleus)”, 传达支撑信息的篇章单元称作“卫星 (satellite)”, 从而形成“核-卫星”的“单核”关系。当修辞关系连接的篇章单元无主次之分时, 则称其为“多核”关系。修辞结构理论认为功能语块是最基本的篇章单元并且两个篇章单元之间的语义关系具有开放性和可扩充性, 这为后来基于修辞结构理论的语料库标注研究的多元发展打下基础。

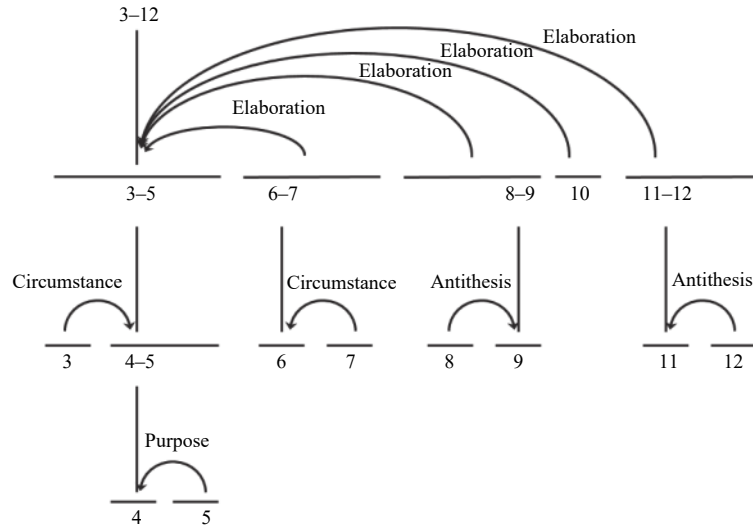


图 3 修辞结构理论篇章结构示例<sup>[1]</sup>

• 篇章词汇化树衔接语法理论与修辞结构理论不同, 该理论来源于词汇化树衔接语法理论 (lexicalized tree-adjoining grammar, LTAG)<sup>[55]</sup>, 即不对文章的整体结构做出任何假设, 只关注局部和词汇触发的篇章关系。因此, 在该理论中, 局部的篇章结构主要由两个篇章单元、两个篇章单元间的篇章关系和能够表示该篇章关系的连接词构成, 随后发展为宾州树库理论。尽管该理论表示的篇章结构可以隐含类似于修辞结构理论的树型结构, 但是主流研究仍然关注于局部的篇章单元间的篇章关系分类和连接词识别等任务, 图 4 是基于此理论构建的篇章语料库 PDTB 中的显式篇章关系示例。

**Connective:** *while*

**Arg1:** Fujitsu Ltd.'s top executive took the unusual step of publicly apologizing for his company's making bids of just one yen for several local government projects,

**Arg2:** computer rival NEC Corp. made a written apology for indulging in the same practice.

**Relation:** TEMPORAL

图 4 篇章词汇化树衔接语法理论示例 (显式篇章关系)

• 分段式篇章表示理论来源于篇章表示理论 (discourse representation theory, DRT)<sup>[5]</sup>, 它将篇章单元间的联系看作是一个图型结构。具体而言, 与修辞结构理论认为的篇章关系只能存在于两个相邻篇章单元间的假设不同, 该理论认为不相邻的两个篇章单元之间也存在远程语义联系。在篇章表示理论 (DRT) 的基础上, 该理论认为在解释篇章结构时, 不仅需要考虑两个篇章单元间的句法关系, 还应该考虑两个篇章单元间的修辞关系, 如解释、叙述等, 如图 5 所示。

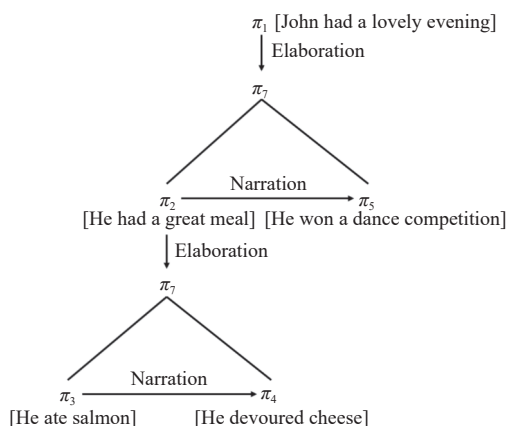


图5 分段式篇章表示理论示例<sup>[49]</sup>

● 跨文档结构理论表示一组相关文本文档的相关性。不同于上文介绍的篇章结构分析理论，跨文档结构理论用于多文档而不是单文档，研究的是多文档之间的篇章结构及其联系，该理论可作为多文档文摘的基础。在修辞结构理论的单文档篇章关系基础上，跨文档结构理论提供了一套跨文档的篇章关系体系，包括4个层次（单词级别，短语级别，段落/句子级别和文档级别）的24种关系。跨文档结构理论认为，不仅文档内部具有一定的篇章结构，文档之间也存在着不同层面的语义联系，如图6所示。

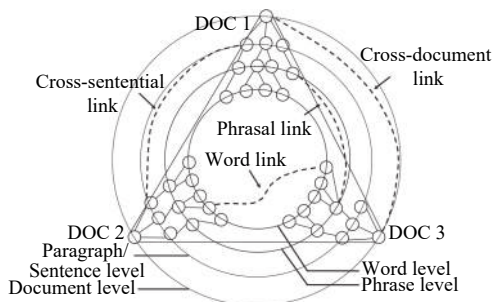


图6 跨文档结构理论示例<sup>[4]</sup>

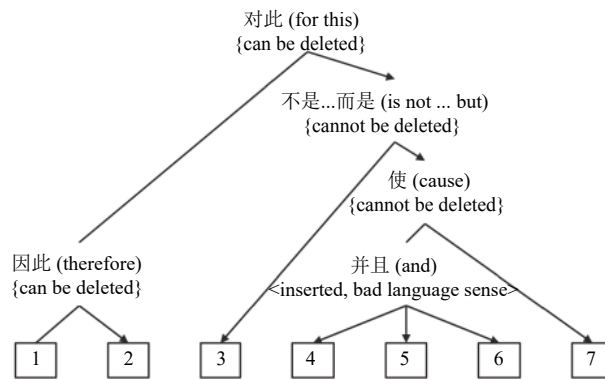
## 2.2 汉语篇章分析理论

● 复句理论和句群理论是中国传统语言学理论。复句理论认为复句由两个或两个以上意义相关、结构上互不作为句子成分的分句组成。在更高的语义层面，句群理论则认为由前后连贯共同表示一个中心意思的几个句子组成的文本块称为句群。因此，不仅组成句群的拥有共同中心思想的句子间存在紧密的逻辑关系，而且各个句子还需要通过语义组合或关联组合的方式组合在一起才能成为一个句群。在复句理论和句群理论中，句子和句子之间的结构与语义联系与英语篇章结构分析理论的篇章结构和关系较为相似。

● 汉语框架语义网理论构建了一系列的语义框架，通过对于句子中的知识框架化，以词元为核心，利用已构建的语义框架映射，表示句子内部词与词之间的关系。在汉语框架语义网理论中，一个句子由词元、核心框架元素和非核心框架元素3部分组成，用于表现句子的语义。并且，框架和框架之间也存在语义联系，如继承关系、总分关系、因果关系和参照关系等。与传统篇章结构分析理论直接构建篇章单元间语义关系不同，该理论中句子之间的语义关系是通过先将句子各部分映射到框架语义网之中，然后通过框架元素之间的关系建立连接。

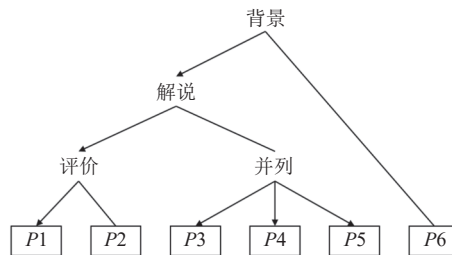
● 基于连接依存树的汉语篇章结构理论则结合 PDTB 体系中连接词驱动策略和 RST 体系中层次化的树型篇章结构，同时结合汉语复句和句群理论中汉语篇章单元的定义与划分方式对汉语篇章结构进行表示。因此，该理论把一个篇章表示为一棵结构树并且指明了篇章单元间的主次和修辞关系。另外，该理论还在两个篇章单元之间通过标识显式或者隐式的连接词暗含其篇章关系。图7是基于连接依存树的汉语篇章结构理论的标注样例。



图 7 基于连接依存树的汉语篇章结构理论示例<sup>[3]</sup>

• 中文篇章级句间语义关系体系针对英语篇章结构分析理论迁移到汉语中存在语义覆盖不完整、关系分类不清、时态关系迁移困难、元素定义困难等问题,该理论体系提出适应于汉语的 5 大类 14 小类 22 种的 3 层篇章关系体系.与篇章词汇化树衔接语法理论类似,该理论主要使用浅层形式表示汉语句间语义关系,具体包括篇章单元、显式篇章关系和隐式篇章关系 3 个部分.

• 汉语宏观篇章结构表示理论则借鉴修辞结构理论<sup>[1]</sup>和宏观结构理论<sup>[56]</sup>构建了以段落为基本篇章单元,依靠段落之间的篇章关系而构成整篇文章的树型篇章结构.图 8 是汉语宏观篇章结构表示理论标注样例,与修辞结构理论一致,该理论也将篇章结构看作是树型结构,并同样指明了篇章单元之间的主次和修辞关系.由于段落之间含有较少的连接词,该理论只提出适应于段落之间的 3 大类 15 小类篇章关系体系,而没有标识类似基于连接依存树的汉语篇章结构理论中的显式或隐式连接词.基于连接依存树的汉语篇章结构理论更侧重于段落内部的子句和句子之间的篇章结构,因此关注于段落之上篇章结构的汉语宏观结构表示理论是对于完整的汉语篇章结构分析理论的有益补充.

图 8 汉语宏观篇章结构表示理论示例<sup>[54]</sup>

### 2.3 小结

综上所述,英语的篇章结构分析理论发展较为完善,多数篇章结构分析理论认为篇章结构是存在于句子级别之上的文章各部分之间的结构和关系.各个篇章结构分析理论分别从浅层、树型和图型 3 种形式表示篇章结构,既有关于子句、句子等微观层面的分析理论,也有关注段落、篇章甚至跨篇章等宏观层面的分析理论.而汉语篇章结构分析理论仍然不够完善.从传统篇章语言学理论如复句理论和句群理论到与国际接轨的篇章结构分析理论,如汉语框架语义网理论,基于连接依存树的汉语篇章结构理论,中文篇章级句间语义关系体系和汉语宏观篇章结构分析理论等,大部分研究借鉴英语篇章结构分析理论,尚未形成较为多样和完善的篇章结构分析理论体系.各个篇章结构分析理论在篇章单元颗粒度、篇章关系种类和篇章结构表示形式上的差异如表 1 所示.另外,尽管各个篇章结构分析理论定义的篇章关系种类不同,但是也存在一些相似的关系,如解说关系、并列关系、因果关系等,如表 2 所示.

表1 篇章结构分析理论比较

语种	名称	篇章单元颗粒度	篇章关系种类	篇章结构表示形式
英语	Hobbs模型	子句、句子	12	浅层
	D-LTAG	子句、句子	—	浅层
	RST	子句、句子、段落	4/25	树型
	SDRT	子句、句子	15	图型
	CST	单词、短语、句子/段落、文档	24	图型
汉语	复句理论和句群理论	子句、句子	—	浅层
	汉语框架语义网理论	子句、句子	4	浅层
	中文篇章级句间语义关系体系	子句、句子	5/14/22	浅层
	基于连接依存树的汉语篇章结构理论	子句、句子	4/17	树型
	汉语宏观篇章结构表示理论	段落	3/15	树型

表2 不同篇章结构分析理论的关系集合

篇章结构分析理论	篇章关系集合
Hobbs模型	Enablement, Cause, Evaluation, Functional, Visual, Explanation, Parallel, Generalization, Exemplification, Contrast, Elaboration, Violated Expectation
RST	Evidence, Justify, Concession, Antithesis, Circumstance, Solutionhood, Elaboration, Background, Enablement, Motivation, Volitional Cause, Non-Volitional Cause, Volitional Result, Non-Volitional Result, Purpose, Condition, Otherwise, Interpretation, Evaluation, Restatement, Summary, Sequence, Joint, Contrast, Reason
SDRT	Elaboration, Entity-Elaboration, Comment, Flashback, Background, Goal, Explanation, Attribution, Narration, Contrast, Result, Parallel, Continuation, Alternation, Conditional
CST	Identity, Equivalence, Translation, Subsumption, Contradiction, Historical background, Cross-reference, Citation, Modality, Attribution, Summary, Follow-up, Elaboration, Indirect speech, Refinement, Agreement, Judgement, Fulfilment, Description, Reader profile, Contrast, Parallel, Generalization, Change of perspective
汉语框架语义网理论	继承, 总分, 因果, 参照
中文篇章级句间语义关系体系	同步, 先序, 后序, 充分条件, 必要条件, 无条件, 形式条件, 同向对比, 反向对比, 间接对比, 原因, 结果, 间接因果, 目的, 解释说明, 实例, 例外, 泛化, 递进, 平行, 相容选择, 互斥选择
基于连接依存树的汉语篇章结构理论	因果, 推断, 假设, 目的, 条件, 背景, 并列, 顺承, 递进, 选择, 对比, 转折, 让步, 解说, 总分, 例证, 评价
汉语宏观篇章结构表示理论	背景, 解说, 果因, 因果, 行为目的, 总结, 评价, 顺承, 对比, 并列, 陈述举例, 目的行为, 递进, 补充, 举例陈述

### 3 英语篇章结构分析研究

随着以上篇章分析理论的提出, 大量的篇章语料库被构建, 如表3所示。例如基于修辞结构理论的 RST-DT<sup>[29]</sup>, Instruction-DT<sup>[57]</sup>, GUM<sup>[58]</sup>和基于远程监督构建的大规模篇章语料库 MEGA-DT<sup>[59]</sup>; 基于分段式篇章表示理论的 STAC<sup>[60]</sup>和 Molweni<sup>[61]</sup>; 基于篇章词汇化树衔接理论的 PDTB<sup>[30]</sup>, Edina-DR<sup>[62]</sup>和 TED-MDB<sup>[63]</sup>等。本文以具有代表性的基于层次化篇章关系的语料库 (RST-DT<sup>[29]</sup>) 和基于扁平化篇章关系的语料库 (PDTB<sup>[30]</sup>) 的相关工作进行重点介绍。

表3 代表性英语篇章结构分析语料库

篇章结构分析理论	代表性语料库
基于RST	RST-DT, Instruction-DT, GUM, MEGA-DT
基于D-LTAG	PDTB, Edina-DR, TED-MDB
基于SDRT	STAC, Molweni

### 3.1 层次化篇章关系语料库 (RST-DT)

美国南加利福尼亚大学标注的 RST-DT<sup>[29]</sup>由 LDC 于 2002 年正式发布. RST-DT 选用宾州树库文章作为标注对象, 基于修辞结构理论构建篇章结构树. 该理论将一篇文章的结构表示为树型结构, 并以此标注篇章单元间的主次和修辞关系, 因此具有层次化的篇章关系. 特别地, RST-DT 对基本篇章单元进行严格定义, 明确了英语中基本篇章单元的划分标准. RST-DT 包含 385 篇已标注文章, 在篇章关系的标注上继承并发展了修辞结构理论原有的 25 类篇章关系, 共标注了 16 组 78 种关系 (包含 53 种单核关系和 25 种多核关系), 每组都具有相同的修辞功能.

在 RST-DT 上的篇章结构分析研究主要包含两个任务: 基本篇章单元分割和篇章结构树构建. 由于基本篇章单元分割任务已经取得较优性能<sup>[64-66]</sup>, 其  $F1$  值达到 96.30%, 因此, 目前的研究重心倾向于篇章结构树构建任务.

传统的篇章结构树构建的研究工作主要分为机器学习方法以及神经网络方法. 最近, 利用多任务或者跨语言方法和无监督或者半监督方法进行篇章结构分析的研究逐渐流行起来. 本文首先介绍传统机器学习方法和神经网络方法, 然后介绍最近 3 年流行的多任务或者跨语言方法和无监督或者半监督方法.

#### 3.1.1 机器学习方法

根据篇章结构树构建方法的不同, 使用机器学习模型的研究工作主要分为自下而上、基于 CYK (Cocke-Younger-Kasami) 和基于转移 3 种方法.

- 自下而上方法. Hernault 等人<sup>[67]</sup>使用支持向量机 (support vector machine, SVM) 作为分类器抽取特征后, 第一次实现了完整的篇章结构分析器, 该分析器通过判断篇章单元的合并位置贪心地自下而上完成篇章结构树的构建. Feng 等人<sup>[68]</sup>在 Joty 等人<sup>[69]</sup>的研究基础上, 使用条件随机场 (conditional random field, CRF) 模型判断篇章单元的合并位置, 并且分别使用两个模型在句内和句间构建篇章结构树. 在采用大量的特征后, Feng 等人<sup>[68]</sup>使用局部链结构来获得最大概率的合并位置从而自下而上的构建篇章结构树.

- 基于 CYK 方法. Joty 等人<sup>[70]</sup>使用动态条件随机场 (dynamic conditional random field, DCRF) 模型计算篇章单元的合并位置, 通过全局最优算法 (CYK 方法) 构建了句内的篇章结构分析器. 在此基础上, Joty 等人<sup>[69]</sup>将篇章划分为句内和句间两个层次, 分别利用 DCRF 构建了篇章结构分析器.

- 基于转移方法. Ji 等人<sup>[71]</sup>使用基于转移的方法通过特征投影完成其向量表示, 然后利用 SVM 模型作为分类器构建了完整的篇章结构分析器. 相比较基于 CYK 方法<sup>[69]</sup>, Ji 等人<sup>[71]</sup>通过更好的篇章单元表示在主次识别和关系识别上都有了进一步的提升. Heilman 等人<sup>[72]</sup>使用逻辑斯蒂回归模型进行多分类实现了一个快速的可以应用到实践中的篇章结构分析器, 该方法使用基于转移的方法构建篇章结构树. Wang 等人<sup>[73]</sup>使用支持向量机 (SVM) 模型作为转移方法的局部模型构建篇章结构树, 并且在总结之前的工作基础上, 使用包括语义和结构等多重特征作为模型的输入. 特别的, 该方法在针对篇章关系识别时, 在句内、句间、段间分别构建分类器进行篇章关系识别, 从而取得了更优的性能.

#### 3.1.2 神经网络方法

随着神经网络的发展, 在篇章结构分析领域, 使用神经网络模型作为局部模型的方法逐渐成为主流. 其篇章结构树的构建方法分为基于 CYK、基于转移和自上而下 3 种方法.

- 基于 CYK 方法. Li 等人<sup>[74]</sup>使用递归神经网络作为局部模型对于子句、句子和整个篇章进行联合建模, 构建了篇章结构分析器. 尽管该方法没有取得比机器学习方法更优的性能, 但是其成为将神经网络模型应用到篇章结构分析中的一次成功尝试. Li 等人<sup>[75]</sup>使用层次化的 Bi-LSTM 模型进行篇章单元的向量表示, 并通过基于张量匹配网络判断两个篇章单元间的语义联系, 最后通过全局最优算法进行篇章结构树的构建.

- 基于转移方法. Jia 等人<sup>[76]</sup>在使用 Bi-LSTM 进行篇章单元的向量表示的基础上, 通过记忆网络后使用基于转移的方法 (移进规约算法) 进行篇章结构树构建, 实验结果显示该方法由于增加了记忆网络模块从而对于长文档的篇章结构树构建更加有效. Yu 等人<sup>[77]</sup>使用基于转移的方法 (Bi-affine), 通过抽取隐式句法特征作为输入构建篇章结构分析器. Mabona 等人<sup>[78]</sup>提出一个文档级 RNN 句法器 (document-level RNN grammar, RNRG) 用于篇章结构分析. 尽管 Mabona 等人<sup>[78]</sup>声称该模型是首次使用生成式方法进行篇章结构分析, 但是该方法被认为是基于转移的篇章结构分析方法. Zhang 等人<sup>[79]</sup>提出一个基于递归信息流动门控的篇章结构分析器, 该分析器使用 Bi-LSTM



表示篇章单元的语义, 然后定义一个门控机制来控制信息流动增强篇章单元间的语义交互, 最后使用基于转移的方法构建篇章结构树. Li 等人<sup>[80]</sup>使用基于转移的移进规约算法构建篇章结构树. 特别的, 其提出的篇章结构分析器针对篇章结构树的多分支问题, 在传统的基于移进规约的方法上, 额外增加了对于多分支的合并行为以及多核心类别的预测类别, 从而增强对于此类型结构的识别性能. 在通过 Bi-affine 模型解码后, 在多并列的关系识别上性能有所提升.

• 自上而下方法. Lin 等人<sup>[65]</sup>和 Liu 等人<sup>[81]</sup>分别使用不同的指针网络自上而下构建了句内的篇章结构分析器, 不过并没有在完整的文章上进行篇章结构分析. Kobayashi 等人<sup>[82]</sup>在使用 Bi-LSTM 表示篇章单元语义后通过判断篇章单元序列的切分位置自上而下的构建篇章结构树. 特别的, 该方法在流行的两段式(句内和句间)篇章结构树构建的基础上, 进一步地将篇章结构分析划分为句内、句间、段间 3 个层次, 取得了更优的效果. Koto 等人<sup>[83]</sup>通过自上而下的动态先知方法序列化构建篇章结构树. 在篇章结构树构建任务上, 通过比 RST-Parseval 方法更为严格的原始的 Parseval 评估方法进行评估, 该模型取得 73.1% 的 F1 值, 接近于人类的 78.7% 的 F1 值. Nguyen 等人<sup>[66]</sup>通过预训练模型 XLNet 编码篇章单元, 将基本篇章单元分割和篇章结构树构建联合建模, 使用自上而下方法构建篇章结构树, 在使用 RST-Parseval 评估方法评估后, 该模型在篇章结构树构建任务中取得的性能(87.6% 的 F1 值)已经接近人类水平(88.7% 的 F1 值).

综上所述, 使用神经网络模型的工作主要聚焦在增强篇章单元语义表示、增强篇章单元间的语义交互和更优的篇章结构树构建方法 3 个方面, 如表 4 所示. 在篇章单元语义表示上, 早期的研究工作一般使用 LSTM 及其变种方式进行编码, 最近的研究则倾向于利用预训练模型获取更好的向量表示. 而且, 多数研究使用各种类型的匹配网络捕获篇章单元间的语义交互. 在篇章结构树构建方法上, 早期跟随机器学习方法的研究工作使用基于转移的方法和基于全局最优算法(CYK)构建篇章结构树, 在最近的研究工作中, 自上而下的构建篇章结构树方法逐渐成为主流.

表 4 基于神经网络的篇章结构分析方法比较

模型	篇章单元语义表示	篇章单元语义交互	篇章结构树构建
Li 等人 (2016) <sup>[75]</sup>	双层Bi-LSTM+Attention	基于张量匹配网络+特征	基于CYK
Jia 等人 (2018) <sup>[76]</sup>	Bi-LSTM+特征	记忆网络匹配	基于转移
Yu 等人 (2018) <sup>[77]</sup>	双层Bi-LSTM+Attention+特征	双仿射网络	基于转移
Mabona 等人 (2019) <sup>[78]</sup>	Tree-LSTM+特征	Stack-LSTM+MLP+柱搜索	基于转移
Li 等人 (2020) <sup>[80]</sup>	Bi-LSTM	双仿射网络+特征	基于转移
Lin 等人 (2019) <sup>[65]</sup>	ELMO+BiGRU	指针网络	自上而下
Kobayashi 等人 (2020) <sup>[82]</sup>	Bi-LSTM+Selective Gate	MLP	自上而下
Koto 等人 (2021) <sup>[83]</sup>	BERT+Bi-LSTM+特征	MLP+动态先知	自上而下
Nguyen 等人 (2021) <sup>[66]</sup>	XLNet+Bi-LSTM	指针网络+柱搜索	自上而下

### 3.1.3 基于多任务或跨语言方法

由于最近篇章结构分析工作重心逐渐从单一语言、单一任务向多语言、多任务扩展, 本节将重点介绍利用多任务或跨语言辅助篇章结构分析的相关工作. Braud 等人<sup>[84]</sup>使用多视图多任务方法, 联合学习成分树与依存树生成任务, 并将其他额外标注信息作为辅助任务, 通过构建多层的 LSTM 模型进行篇章结构分析. 该方法采用序列到树的转换方法构建篇章结构树. Braud 等人<sup>[85]</sup>尝试在基于修辞结构理论构建的 6 种不同语言的语料库上进行篇章结构分析. 该方法借鉴句法分析器<sup>[86]</sup>的相关工作构建了篇章结构分析器, 使用基于转移的方法进行篇章结构树的构建. Liu 等人<sup>[87]</sup>探寻将多语言统一处理的方法构建多语言篇章结构分析器. 一方面, 他们尝试使用 XLM-RoBERTa 模型统一多语言的篇章单元的语义向量表示进行多语言的篇章结构分析. 另一方面, 他们尝试使用机器翻译将多语言统一到特定某一语言内, 从而进行多语言的篇章结构分析. 该方法采用指针网络自上而下构建篇章结构树. Guz 等人<sup>[88]</sup>尝试使用同指任务帮助篇章结构分析任务. 该方法通过基于转移的方法进行篇章结构树的构建. 由于使用更为强大的 SpanBERT 预训练模型, 该方法取得了较好的效果.

### 3.1.4 基于半监督或无监督方法

由于篇章语料标注困难、标注工作量大等问题,最近3年的一些研究尝试使用大规模半监督或者无监督的语料进行篇章结构分析。Kobayashi 等人<sup>[89]</sup>将篇章结构分析划分为句内、句间、段间3个层次,使用度量两个篇章单元间的语义相似度的无监督方法,通过全局最优算法(CYK)构建篇章结构树。他们尝试分别以篇章单元的分割点和合并点作为构建篇章结构树的依据,实验结果显示,使用合并篇章单元的方法能够更有效地构建篇章结构树。Nishida 等人<sup>[90]</sup>使用维特比 EM 算法通过句法知识进行无监督学习篇章结构,该方法使用全局最优算法(CYK)构建篇章结构树。Huber 等人<sup>[91]</sup>使用树型自动编码器进行无监督学习篇章结构,并自上而下的构建篇章结构树。Kobayashi 等人<sup>[92]</sup>使用教师-学生框架的半监督方法,以 Yu 等人<sup>[77]</sup>和 Wang 等人<sup>[73]</sup>的篇章结构分析器作为教师分析器,以 Kobayashi 等人<sup>[82]</sup>的模型作为学生分析器,在 CNN 语料库上进行无标签学习,该方法同样使用流行的自上而下的方法构建篇章结构树。Xiao 等人<sup>[93]</sup>通过自动摘要任务辅助篇章结构树的构建。他们首先在 CNNDM/NYT 语料库训练摘要模型,获取其网络内部的注意力得分作为篇章单元间的语义联系密切程度的表示,然后使用全局优化方法 CLE (Chu-Liu-Edmonds) 算法构建篇章结构树。

### 3.1.5 小结

综上所述,最近3年在 RST-DT 中的研究重点发生了转移。早期的篇章结构分析方法研究主要侧重在3个方面:(1)增强篇章单元的语义表示;(2)增强篇章单元间的语义交互;(3)更优的篇章结构树构建方法。最近3年的研究则主要扩展到更多样化的训练语料和更多样化的学习目标两个方面。表5是 RST-DT 上各个模型的性能表现。

表5 各个模型性能的比较 (RST-DT) (%)

类别	方法	模型	RST-Paraseval			Paraseval			
			S	N	R	S	N	R	F
机器学习	自下而上	Hernault等人(2010) <sup>[67]</sup>	74.7	59.0	44.3	—	—	—	—
		Feng等人(2014) <sup>[68]</sup>	85.7	71.0	58.2	68.6	55.9	45.8	44.6
	基于CYK	Joty等人(2013) <sup>[69]</sup>	82.7	68.4	55.7	—	—	—	—
		Ji等人(2014) <sup>[71]</sup>	82.1	71.1	61.6	64.1	54.2	46.8	46.3
	基于转移	Heilman等人(2015) <sup>[72]</sup>	83.5	68.1	55.1	—	—	—	—
		Wang等人(2017) <sup>[73]</sup>	86.0	72.4	59.7	—	—	—	—
神经网络	基于CYK	Li等人(2014) <sup>[74]</sup>	84.0	70.8	58.6	—	—	—	—
		Li等人(2016) <sup>[75]</sup>	85.8	71.1	58.9	64.5	54	38.1	36.6
	基于转移	Yu等人(2018) <sup>*[77]</sup>	85.5	73.1	60.2	71.4	60.3	49.2	48.1
		Mabona等人(2019) <sup>[78]</sup>	—	—	—	67.1	57.4	45.5	45.0
	自上而下	Li等人(2020) <sup>*[80]</sup>	<b>88.4</b>	75.8	63.1	<b>76.5</b>	<b>65.3</b>	<b>53.9</b>	—
		Kobayashi等人(2020) <sup>*[82]</sup>	87.0	74.6	60.0	—	—	—	—
多任务跨语言	—	Koto等人(2021) <sup>*[83]</sup>	—	—	—	73.1	62.3	51.5	50.3
		Nguyen等人(2021) <sup>*[66]</sup>	87.6	76.0	61.8	74.3	64.3	51.6	<b>50.2</b>
		Braud等人(2016) <sup>[84]</sup>	—	—	—	59.5	47.2	34.7	34.3
		Braud等人(2017) <sup>[85]</sup>	—	—	—	62.7	54.5	45.5	45.1
半监督无监督	—	Liu等人(2020) <sup>*[87]</sup>	87.8	75.4	63.5	—	—	—	—
		Guz等人(2020) <sup>*[88]</sup>	88.1	<b>76.1</b>	<b>63.6</b>	—	—	—	—
		Kobayashi等人(2019) <sup>[89]</sup>	81.1	—	—	—	—	—	—
		Nishida等人(2021) <sup>[90]</sup>	84.6	—	—	68.6	—	—	—
		Kabayashi等人(2021) <sup>[92]</sup>	86.8	74.7	62.5	—	—	—	—
—	—	Xia等人(2021) <sup>[93]</sup>	76.7	—	—	—	—	—	
—	—	Huber等人(2021) <sup>[91]</sup>	71.3	—	—	—	—	—	
—	—	Human agreement	88.7	77.7	65.8	78.7	66.8	57.1	55.0

注: \*表示其使用了预训练语言模型

从表 5 中有 3 点发现. 首先, 使用预训练模型 (ELMO, BERT, XLNet, SpanBERT 等) 表示篇章单元语义的方法的性能要显著优于未使用预训练模型的方法, 并且使用的预训练模型越先进, 其性能表现越好, 这表明通过预训练模型更好地理解篇章单元的语义有助于篇章结构分析任务. 其次, 使用更加精细化篇章结构树构建步骤也能够取得更优的性能, 这表明篇章结构存在层次化的特点. 最后, 在半监督无监督工作中, 监督信号的强弱与篇章结构树构建性能呈现正相关特性, 这表明篇章结构与其他语义分析任务联系密切.

### 3.2 扁平化篇章关系语料库 (PDTB)

PDTB 是一个基于 D-LTAG 理论的侧重于扁平化篇章关系语料库, 其文章全部来源于华尔街日报. 该语料库并不关注篇章整体的结构表示, 并没有构建一个层次化的篇章结构, 因此为扁平化的篇章关系语料库. 目前, 它流行的是 PDTB 2.0 版本<sup>[30]</sup>, 最新的公开语料库是 PDTB 3.0 版本<sup>[94]</sup>. 在 PDTB 中, 根据标注的内容, 主要有 4 个任务: 连接词检测 (connective detection), 论元识别 (argument labeling), 篇章关系识别 (discourse relation recognition) 和属性识别 (attribute labeling). 目前工作主要聚焦在较为困难和重要的篇章关系识别中.

PDTB 2.0 中的篇章关系根据在文本中是否包含连接词分为两种类型. 第 1 种是拥有显式连接词的显式篇章关系, 第 2 种是两个篇章单元之间不包含显式连接词的隐式篇章关系. 具体而言, PDTB 2.0 共包含 40600 条已标注的篇章关系, 其中主要包括 18439 条显式篇章关系和 16224 条隐式篇章关系. 其篇章关系被分为 4 大类并可以进一步划分为 16 小类和 23 个亚型. 在无法判定隐式关系的情况下, PDTB 2.0 语料库还有其他 3 种不太常见的类型, 包括 AltLex, EntRel 和 NoRel.

在 PDTB 2.0 中, 由于显式篇章关系识别已经取得非常高的性能<sup>[95,96]</sup>, 达到 94.15% 的  $F1$  值, 超过人工标注的 94%, 因此目前的研究工作主要专注于隐式篇章关系识别任务. 根据研究重心的不同, 其相关工作可以被划分为 3 大类: 增强篇章单元语义表示, 增强篇章单元间的语义交互, 联合学习隐式篇章关系识别与其他任务. 由于在 PDTB 2.0 中的研究工作对于语料库的训练集、验证集、测试集划分不同, 为了便于比较各模型的性能表现情况, 下面的研究工作均在官方推荐<sup>[30]</sup>, 并且最常用的数据集划分 (PDTB-Ji) 上进行的实验.

#### 3.2.1 增强篇章单元语义表示

早期的工作主要探索使用神经网络增强篇章单元的语义表示. Zhang 等人<sup>[97]</sup>提出浅层卷积神经网络 (SCNN) 进行篇章关系识别. 该方法通过对词嵌入矩阵进行最大、最小、平均卷积运算得到篇章单元的表示, 然后将其拼接表示一个篇章单元对. Ji 等人<sup>[98]</sup>通过实体增强篇章单元表示, 并利用递归神经网络进行篇章关系识别. Qin 等人<sup>[99]</sup>使用卷积神经网络表示篇章单元语义并直接拼接两个篇章单元的向量表示, 然后通过协作门控神经网络增强篇章单元的语义表示. Bai 等人<sup>[100]</sup>提出组合字符、子词和 ELMO 的向量表示能够有效增强篇章单元的表示.

#### 3.2.2 增强篇章单元间语义交互

也有一批工作侧重于增强篇章单元间的语义交互. Liu 等人<sup>[101]</sup>, Lan 等人<sup>[102]</sup>, Guo 等人<sup>[103]</sup>和 Ruan 等人<sup>[104]</sup>分别提出多种注意力机制增强篇章单元间的语义交互. Liu 等人<sup>[101]</sup>提出使用多层注意力机制捕获两个篇章单元之间的语义交互. Lan 等人<sup>[102]</sup>不仅在两个篇章单元之间使用交互注意力网络, 还在不同的任务之间使用门控注意力网络捕获不同任务之间的相互影响. Guo 等人<sup>[103]</sup>使用 pair-wise 矩阵获得注意力交互后, 使用张量网络进行篇章关系分类. Ruan 等人<sup>[104]</sup>提出交互注意力的同时, 使用双层注意力权重增强两个篇章单元间的语义交互. 另外, 也有一些工作<sup>[105,106]</sup>通过构建更好的上下文用于增强篇章单元间的语义交互. Guo 等人<sup>[105]</sup>提出一种知识增强的注意力网络, 该网络通过引入外部知识来增强篇章单元间的语义交互. Zhang 等人<sup>[106]</sup>通过构建一个上下文追踪网络, 通过图模型进行篇章关系识别, 取得了较好的性能.

#### 3.2.3 联合学习建模篇章关系

为了更好地理解两个篇章单元间的关系, 一些工作采用将隐式篇章关系与其他能够更明显表示篇章单元间关系的任务相结合进行联合学习. Lan 等人<sup>[102]</sup>构建了一个基于注意力的篇章关系识别模型, 通过门控交互机制同时学习隐式篇章关系和显式篇章关系两个任务. Bai 等人<sup>[100]</sup>, Shi 等人<sup>[107]</sup>则选择更加直接的连接词预测与隐式篇章关系识别进行联合学习. 两者不同之处在于 Bai 等人<sup>[100]</sup>通过连接词分类任务辅助隐式篇章关系识别, 而 Shi 等人<sup>[107]</sup>

通过生成方法在原句中插入连接词的形式辅助隐式篇章关系识别. 另外, 通过对于篇章关系进行标签嵌入也能够增强模型对于篇章关系的语义理解. Nguyen 等人<sup>[108]</sup>通过保持篇章关系与连接词的嵌入一致性理解篇章关系语义. He 等人<sup>[109]</sup>提出通过 TransS 将两个篇章单元与篇章关系同时映射到低维空间中, 从而更好地理解篇章关系语义.

### 3.2.4 小结

综上所述, 在扁平化篇章关系语料库 PDTB 2.0 中的研究工作与在层次化篇章关系语料库 RST-DT 中的研究工作重点大致相同. 由于 PDTB 不需要进行结构树的构建, 因此相关工作主要集中在增强篇章单元的语义表示, 增强篇章单元间的语义交互以及与其他任务联合学习 3 个方面, 其研究趋势也从理解两个篇章单元的语义向理解篇章关系本身的语义等更深层的理解语义方向发展, 如图 9 所示.

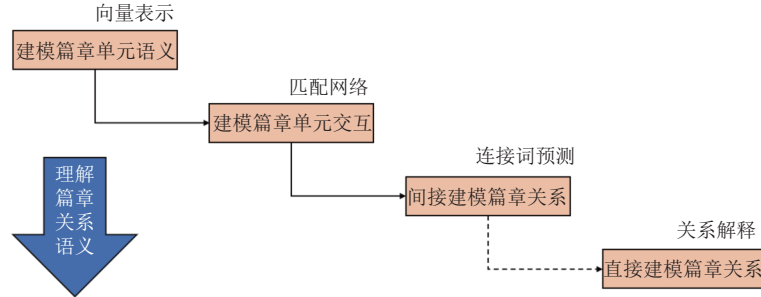


图 9 PDTB 中的研究重点变化情况

在常用的数据集划分 (PDTB-Ji) 下各个模型的性能表现如表 6 所示. 从表 6 中有以下两点发现. 首先, 即使是取得最优性能的模型, 其在 4 大类篇章关系识别的效果仍然不够理想, 最高性能只达到 53.11% 的  $F1$  值, 这说明隐式篇章关系识别仍然具有挑战性. 其次, 在 4 类篇章关系类别上取得最优值的模型并不是同一个模型, 这说明不同的模型在识别不同类别的篇章关系上有不同的表现.

表 6 各个模型性能的比较 (PDTB) (%)

类别	模型	Comp.	Cont.	Exp.	Temp.	4-way
增强篇章单元语义表示	Zhang 等人 (2015) <sup>[97]</sup>	32.22	52.04	69.59	30.54	—
	Ji 等人 (2015) <sup>[98]</sup>	35.93	52.78	—	27.63	—
	Qin 等人 (2016) <sup>[99]</sup>	41.55	57.32	71.50	35.43	—
增强篇章单元间语义交互	Liu 等人 (2016) <sup>[101]</sup>	39.86	53.69	69.71	37.61	44.95
	Lan 等人 (2017) <sup>[102]</sup>	40.73	58.96	72.47	38.50	47.80
	Guo 等人 (2018) <sup>[103]</sup>	40.35	56.81	72.11	38.65	47.59
	Guo 等人 (2020) <sup>[105]</sup>	43.92	57.67	73.45	36.33	47.90
	Ruan 等人 (2020) <sup>[104]</sup>	46.75	59.56	<b>75.83</b>	39.35	—
	Zhang 等人 (2021) <sup>[106]</sup>	46.86	55.63	73.71	<b>45.90</b>	<b>53.11</b>
联合学习建模篇章关系	Bai 等人 (2018) <sup>[100]</sup>	47.85	54.47	70.60	36.87	51.06
	Shi 等人 (2019) <sup>[107]</sup>	41.83	<b>62.07</b>	69.58	35.72	46.40
	Nguyen 等人 (2019) <sup>[108]</sup>	<b>48.44</b>	56.84	73.66	38.60	53.00
	He 等人 (2020) <sup>[109]</sup>	47.98	55.62	69.37	38.94	51.24

## 4 汉语篇章结构分析研究

近年来随着汉语研究理论逐渐与国际研究同步, 许多基于汉语篇章结构分析理论的篇章语料库被构建, 包括标注新闻语料的 HIT-CDTB<sup>[33]</sup>, SUDA-CDTB<sup>[3]</sup>, LDC-CDTB<sup>[34]</sup>以及 MCDTB<sup>[31]</sup>等, 也有标注口语化语料的 TED-CDB<sup>[110]</sup>和标注学术论文摘要语料的 Sci-CDTB<sup>[111]</sup>. 各语料库概况如表 7 所示.



表7 代表性汉语篇章结构分析语料库

代表性语料库	层次化结构	关系类别	标注规模(篇)	标注风格	体裁
HIT-CDTB	无	22	525	PDTB	OntoNotes 4.0 (网络新闻)
LDC-CDTB	无	11	164	PDTB	CTB (新闻报道)
TED-CDB	无	30	72	PDTB	TED (演讲)
SUDA-CDTB	有	17	500	RST, PDTB	CTB (新闻报道)
MCDTB	有	15	720	RST	CTB (新闻报道)
Sci-CDTB	有	26	108	RST	JCIP (学术论文)

从表7中可以看出不同语料库的差异主要是在是否标注层次化结构, 篇章关系类别数量, 标注规模, 标注风格以及体裁等5个方面。具体而言, HIT-CDTB, LDC-CDTB 和 TED-CDB 均是 PDTB 风格的标注方式, 该方式标注的篇章关系不包含层次化的篇章结构。而 SUDA-CDTB 和 MCDTB 是 RST 风格的标注方式, 把篇章结构标注为树型并且额外标注了篇章单元间的主次和修辞关系。Sci-CDTB 依据 RST-DT 的规则划分基本篇章单元, 但是其篇章结构则采用依存结构树而不是成分结构树的标注方式。就体裁而言, HIT-CDTB, LDC-CDTB, SUDA-CDTB 和 MCDTB 等均采用结构规范, 条理清晰的新闻作为语料来源, 而最新标注的 TED-CDB 选取更加口语化的 TED 演讲作为语料来源, 只标注 72 篇文章。但是由于体裁的不同, 文章长度为一篇演讲, 因此每篇包含的篇章关系较多。Sci-CDTB 标注了来源于《中文信息学报》的学术论文摘要的篇章结构与关系。本文以侧重于具有代表性的基于层次化篇章关系语料库 (SUDA-CDTB 和 MCDTB) 和侧重于扁平化篇章关系语料库 (HIT-CDTB 和 LDC-CDTB) 的相关工作作为重点进行介绍。

#### 4.1 层次化篇章关系语料库

##### 4.1.1 SUDA-CDTB

Li 等人<sup>[2,3]</sup>在基于连接依存树的汉语篇章结构理论指导下, 标注了 CTB 6.0 中的 500 篇文章的篇章修辞结构, 构建了汉语篇章树库 (SUDA-CDTB)。它将文章中每个段落标注为一棵连接依存树, 标注内容包括基本篇章单元、连接词 (显式关系和隐式关系)、层次化的篇章结构、篇章单元主次和篇章单元间修辞关系 (4 大类共 17 小类)。目前基于 SUDA-CDTB 的研究主要聚焦在篇章结构树构建、主次识别和关系识别 3 个子任务。

Kong 等人<sup>[112]</sup>采用流水线的方式利用最大熵模型实现包括子句识别、连接词识别与分类、隐式篇章关系识别和篇章单元主次识别等任务, 最终输出完整的篇章结构树。Jia 等人<sup>[113]</sup>利用 stack-LSTM 作为基础模型并提出将篇章结构分析划分为句内和句间两个层次, 通过移进规约算法获得一个从文本序列到完整的树型篇章结构。孙成等人<sup>[114]</sup>构建了一个完整的基于转移的篇章结构树的生成框架, 它首先利用卷积神经网络和抽取的特征表示篇章单元, 然后使用 SPINN 作为局部判别模型, 通过移进规约算法构建篇章结构树。Wang 等人<sup>[115,116]</sup>提出使用 Bi-LSTM 和注意力机制表示篇章单元, 并利用张量网络捕获相邻的两个篇章单元的语义交互。它通过基于转移的方法在句内和句间层次化构建篇章结构树。Zhang 等人<sup>[117]</sup>提出以 GloVe 作为词向量表示, 通过指针网络自上而下构建篇章结构树。在此基础上, Zhang 等人<sup>[118]</sup>使用更为先进的 XLNet 作为词向量表示, 在考虑局部最优的自上而下构建篇章结构树的同时, 通过对抗学习判别生成篇章结构树和标准篇章结构树从而考虑全局优化信息, 取得了更优的效果。表 8 为 SUDA-CDTB 上篇章结构树构建任务中各个模型的性能表现。

也有一些工作专注于篇章主次识别和篇章关系识别任务。Xu 等人<sup>[119,120]</sup>提出 TMN (text matching network) 模型。该模型首先使用 Bi-LSTM 进行篇章单元的语义表示, 然后利用由 3 个子模块 (余弦相似度, 双线性 and 单层非线性) 构成的匹配网络捕获篇章单元间的语义交互, 从而进行篇章主次识别。王体爽等人<sup>[121]</sup>利用 Bi-LSTM 表示篇章单元的语义, 然后使用一个多层结构的门控记忆网络捕获篇章单元间的语义交互, 从而进行篇章主次识别。

徐昇等人<sup>[122]</sup>提出使用自注意力机制用于表示篇章单元语义, 然后利用交互注意力和记忆注意力增强篇章单元间语义交互进行篇章关系识别。Xu 等人<sup>[123]</sup>在使用门控卷积网络 (gated convolutional network, GCN) 表示篇章单元的同时, 提出一个简化主题模型用于捕获两个篇章单元的语义交互, 并最终使用张量分解模型进一步增强语义交互进行篇章关系识别。



表 8 各个模型性能的比较 (SUDA-CDTB) (%)<sup>[118]</sup>

模型	Structure	Nuclearity	Relation	Full
Sun等人 (2018) <sup>[112]</sup>	84.8	55.8	52.1	47.7
Zhang等人 (2020) <sup>[117]</sup>	85.2 (84.0)	57.3 (59.0)	53.3 (54.2)	45.7 (47.8)
Zhang等人 (2021) <sup>[118]</sup>	(86.6)	(65.0)	(62.1)	(55.4)

注: 括号内表示使用更加严格的评估指标

#### 4.1.2 MCDTB

由于基于连接依存树的篇章结构理论的 SUDA-CDTB 标注侧重于段落内部句子之间的篇章关系, 忽略了更高层次的段落之间的篇章关系, 因此 Jiang 等人<sup>[31,32]</sup>在 CTB 8.0<sup>[124]</sup>中选取了 720 篇文章, 根据汉语宏观篇章表示理论<sup>[53]</sup>标注其宏观篇章结构, 即段落之上的篇章结构, 其中部分标注文章与 SUDA-CDTB 标注的文章重合. 与 RST 风格的语料库类似, 在该语料库中, 一篇文章被标注为一棵篇章结构树. 考虑到段间的篇章关系与句间的篇章关系分布不同, 在参考 SUDA-CDTB 的篇章关系标注规范的基础上, MCDTB 提出适应段间关系的 3 大类 15 小类篇章关系标注体系, 并且标注了篇章单元的主次关系. 另外, 它额外标注了段落主题句、文章摘要、篇章单元语用等信息用于其他篇章级任务, 如自动摘要等.

在 MCDTB 中, 最小的篇章单元为段落, 其边界可以自然获得, 因此相关研究工作的主要任务为构建篇章结构树. 然而, 汉语宏观篇章结构分析研究主要专注于 3 个子任务 (篇章结构识别、主次识别和关系识别) 之一或者联合学习其中的两个任务, 最近的研究工作才完成完整的篇章结构分析器的构建.

Jiang 等人<sup>[125]</sup>分别利用结构信息和宏观语义信息构建两个条件随机场模型, 并通过最大概率组合两个模型后识别主次关系, 再利用标签退化方法完成篇章结构识别. 需要注意的是, 该研究工作没有构建完整的篇章结构树. Zhou 等人<sup>[126,127]</sup>在使用词对相似网络计算两个相邻的篇章单元的语义关系同时, 还利用 Bi-LSTM 编码篇章单元与话题之间的语义联系, 并通过最大池化、注意力池化和词对相似网络等多个视图获得篇章单元间的语义交互信息, 最终使用基于转移的方法 (移进规约算法) 构建篇章结构树. Fan 等人<sup>[128]</sup>提出融合全局和局部信息的指针网络模型 (pointer network on global and local information, PNL), 利用全局信息自上而下构建篇章结构树的同时, 充分考虑局部篇章单元间语义交互信息. Jiang 等人<sup>[129]</sup>利用英汉篇章结构的不同倾向, 提出全局和局部反向阅读的方式构建篇章结构树. 值得注意的是, 该工作构建了第一个在 MCDTB 上的完整篇章结构分析器, 包括篇章结构树构建、主次识别和关系识别. Jiang 等人<sup>[130]</sup>通过在段落之上的隐式边界——话题边界帮助模型层次化构建篇章结构树. 表 9 是各个篇章结构分析器的性能比较, 从表 9 中可以看出, 与英语中研究趋势相似, 使用预训练语言模型的篇章结构分析器性能要远优于未使用预训练语言模型的篇章结构分析器的性能.

表 9 各个模型性能的比较 (MCDTB) (%)

模型	Structure	Nuclearity	Relation
Jiang等人 (2018) <sup>[125]</sup>	54.71	—	—
Zhou等人 (2019) <sup>[126]</sup>	56.11	—	—
Fan等人 (2020) <sup>[128]</sup>	58.42	—	—
Jiang等人 (2020)* <sup>[129]</sup>	67.08	56.41	37.09
Jiang等人 (2021)* <sup>[130]</sup>	66.31	55.80	35.39

注: \*表示其使用了预训练语言模型

褚晓敏等人<sup>[131]</sup>认为识别宏观篇章结构中的主次关系对于把握篇章整体结构, 理解文本语义非常重要. 蒋峰等人<sup>[132]</sup>提出基于 Word2Vec 的方法计算两个篇章单元的语义相似度的同时, 使用 LDA 模型作为主题相似度计算方法, 从而在不同维度上计算两个篇章单元之间的语义联系, 并通过最大熵模型识别篇章主次关系. 孙振华等人<sup>[133]</sup>利用 BERT 模型同时编码篇章单元与篇章主题, 捕获篇章单元的语义表示的同时, 也捕获其与篇章主题之间的语义关系, 因此更好的识别篇章主次关系.

周懿等人<sup>[134]</sup>在使用组织结构特征的基础上, 利用 Word2Vec 和 Glove 词向量模型表示篇章单元的局部语义和全局语义, 通过 SVM 模型进行篇章关系分类. Jiang 等人<sup>[135]</sup>提出基于话题和结构门控语义网络的篇章主次和关系联合模型, 它通过话题和结构特征控制篇章单元的语义表示, 从而进行篇章关系识别. Sun 等人<sup>[136]</sup>通过构建词与词, 词与篇章单元间的语义关系后, 利用图卷积网络捕获篇章单元间的语义联系, 再捕获两个篇章单元的语义表示进行篇章关系分类. Jiang 等人<sup>[137]</sup>提出基于多维度 and 层次化建模的篇章关系识别方法. 该方法一方面利用关系大类和关系小类的层次化关系更好的识别关系模型, 另一方面联合建模篇章单元间的修辞、同指和时序 3 个维度的关系, 使得模型对于两个篇章单元间的关系理解更加全面. 表 10 为 MCDTB 上各个模型在篇章关系识别任务上的性能比较.

表 10 各个模型在篇章关系识别的性能比较 (MCDTB) (%)

模型	Sub-class		Class	
	micro-F1	macro-F1	micro-F1	macro-F1
Zhou等人 (2019) <sup>[134]</sup>	54.18	14.92	65.62	50.10
Jiang等人 (2019) <sup>[135]</sup>	56.71	15.68	67.63	57.87
Sun等人 (2020) <sup>[136]</sup>	—	—	70.01	55.38
Jiang等人 (2021) <sup>[137]</sup>	63.27	26.82	72.14	66.87

## 4.2 扁平化篇章关系语料库

除了以上两个侧重于层次化的篇章结构语料库的研究之外, 在侧重扁平化的篇章关系语料库 HIT-CDTB 和 LDC-CDTB 中篇章关系识别任务也是研究热点之一.

### 4.2.1 HIT-CDTB

HIT-CDTB 由哈尔滨工业大学参照 PDTB 方式标注的篇章关系语料库. 该语料库共包含 525 篇文章, 涵盖多个领域, 如新闻、科普文章等. 其标注的篇章关系存在于短语间、子句间、句子间或者句群间, 共有 4 大类和 22 小类.

田文洪等人<sup>[138]</sup>提出多任务的 LSTM 模型用于篇章关系识别. 该模型包含 3 个 LSTM 层, 其中两个 LSTM 层用于篇章单元的语义表示, 分别学习显式篇章关系和隐式篇章关系, 第 3 个 LSTM 层学习两个任务中篇章单元间的语义交互. 唐裕婷等人<sup>[139]</sup>根据篇章关系方向性特点, 通过远程监督的方式自动标注大量显式数据, 从而获得更好的词向量表示, 在 HIT-CDTB 上进行细颗粒度的篇章关系识别. Guo 等人<sup>[140]</sup>提出 DC-BCNN 网络 (dynamic chunk-based max pooling BiLSTM-CNN framework) 进行篇章关系识别. 该模型首先利用 Bi-LSTM 进行篇章单元的语义表示, 然后通过卷积神经网络和动态块最大池化操作提取特征, 最后合并两个篇章单元表示进行篇章关系识别. 由于评价指标不同, 因此在 HIT-CDTB 中各个研究工作的性能表现无法进行直观的比较.

### 4.2.2 LDC-CDTB

LDC-CDTB 由布兰迪斯大学参照 PDTB 方式标注的篇章关系语料库. 不同于原始 PDTB 拥有双层的篇章关系类型标注体系, Zhou 等人<sup>[34]</sup>采用一级篇章关系共 11 类进行标注. 其标注的 164 篇文章来源于更加规范的 CTB 新闻语料.

Qin 等人<sup>[141]</sup>使用卷积网络以典型深度语义匹配模型 (deep structured semantic model, DSSM) 分别编码两个篇章单元, 通过最大池化后拼接两个篇章单元的进行篇章关系分类. Liu 等人<sup>[142]</sup>提出一种新型记忆增强的注意力网络模型, 该模型使用基于注意力的神经网络表示篇章单元, 并设计一个外部记忆网络用于保存关键信息, 从而可以捕获每个篇章关系的聚类结构用于篇章关系识别. Rönnqvist 等人<sup>[143]</sup>提出基于 LSTM 的注意力网络. 该网络没有使用传统的深度语义匹配模型编码两个篇章单元而是利用标识符表示两个篇章单元, 然后使用 LSTM 顺序编码两个篇章单元, 并通过注意力网络获取两个篇章单元的语义交互. 实验结果显示, 该模型取得了更优的性能. 表 11 为 LDC-CDTB 中各模型在篇章关系的识别性能比较.

表 11 各个模型在篇章关系识别的性能比较 (LDC-CDTB) (%)

模型	Acc.
Qin等人 (2016) <sup>[141]</sup>	67.04
Liu等人 (2017) <sup>[142]</sup>	66.02
Rönnqvist等人 (2017) <sup>[143]</sup>	73.01

### 4.3 小结

综上所述,在汉语篇章结构分析研究中,对于微观层面的篇章结构分析研究较为丰富,既有包括完整的篇章结构树构建研究,也有专注于特定任务(如篇章关系识别)的研究,其方法大多与英语研究方法一致。然而,目前汉语宏观篇章结构的研究大多仿照英语的研究以及微观研究的方法,没有充分考虑篇章宏观结构的特点,其研究还处在探索阶段。另外,在宏观篇章结构构建研究中,一般研究重点也侧重于增强篇章单元语义表示、增强篇章单元间的语义交互、更优的篇章结构树构建方法等3个方面。在宏观篇章主次识别任务中,利用话题(主题)信息能够较好地辅助篇章主次关系识别,而在宏观篇章关系分类中,主流研究工作则侧重于增强篇章单元间的语义交互。

## 5 基于篇章结构分析的应用

通过篇章结构分析,能够梳理出文章的脉络以及各部分之间的关系,有助于对于文章整体结构的把握和高层语义的进一步理解。因此,篇章结构分析的结果能够被广泛应用于文本理解和文本生成等自然语言处理下游任务。

### 5.1 文本理解

篇章结构分析的结果包括篇章结构、篇章单元的主次和篇章单元间的修辞关系3个部分。利用这3个部分的信息能够有效帮助理解文本本身的语义,为下游任务提供帮助。

#### 5.1.1 情感分析

在情感分析任务中,篇章结构分析的主次标记最早被应用于情感分析任务<sup>[6]</sup>,而后相关工作也探索利用修辞关系<sup>[7,8]</sup>和利用完整的篇章结构信息<sup>[9]</sup>辅助情感分析。语义取向计算理论<sup>[6]</sup>认为一个篇章的整体情感极性由篇章单元中最重要的部分的情感极性决定。Heerschop等人<sup>[7]</sup>对语义取向计算理论进行了改进,认为情感极性不仅仅取决于篇章单元的主次,还依赖于篇章单元间的修辞关系。而后,Fu等人<sup>[8]</sup>将其理论应用到神经网络方法中,获得了更好的性能。更进一步的,Kraus等人<sup>[9]</sup>提出 Discourse-LSTM 模型将篇章结构树信息融入神经网络中进行情感分析。

#### 5.1.2 信息抽取

篇章结构信息对于信息抽取任务也能够提供帮助,如搜索引擎、句子压缩和论点挖掘等任务。Kuyten等人<sup>[144]</sup>提出第1个基于篇章信息的搜索引擎 DSE,该搜索引擎利用文档的修辞结构树改进用户查询和目标文档之间相关性的度量。DSE 通过衡量候选语句与给定的查询语句之间的修辞关系,不仅可以给出相关的目标文档,还可以在文档中抽取出确切的语句作为查询结果。而且,DSE 能够在一篇包含特定主题的不同意见的文章中准确查找到目标意见。Sporleder等人<sup>[145]</sup>利用主次关系,通过保留“核心”基本篇章单元,去除“卫星”基本篇章单元对句子进行压缩,从而在获得更精简的句子的同时尽可能地保留完整的语义。Hewett等人<sup>[146]</sup>采用 RST 风格和 PDTB 风格的篇章分析器得到篇章特征,并评估自动得到的篇章特征对论点挖掘的贡献程度。实验结果表明加入 PDTB 风格的篇章特征对论点挖掘有促进作用。

#### 5.1.3 其他语义理解任务

另外,篇章结构信息也有助于其他语义理解任务,如阅读理解和文本分类等。Narasimhan等人<sup>[147]</sup>提出了一种将篇章信息融入机器阅读理解的新方法。该方法与传统的使用篇章分析器得到篇章信息的方法不同,它能在模型优化特定任务目标的同时,诱导了句子之间的关系,使模型能够在训练过程中受益于篇章信息,而不依赖于话语结构的显式标注。Gao等人<sup>[148]</sup>提出了一个具有篇章感知能力的隐式推理网络,使用预先训练的篇章分割模型将文档

分割为基本篇章单元,并以弱监督的方式训练模型,以预测每个基本篇章单元是否被对话中的用户反馈所包含。基于切分得到的基本篇章单元和学习到的其隐含表示进行阅读理解。Ji 等人<sup>[149]</sup>利用生成的篇章结构树更好的表示文本的语义,从而辅助模型进行文本分类。

## 5.2 文本生成

早期篇章结构理论是为指导文本生成而设计,但是受限于当时的技术手段,更多的研究重点转移到文本理解任务中。最近,随着机器学习和深度学习技术的发展,越来越多的研究者重新尝试利用篇章结构信息进行文本生成任务,如自动摘要、机器翻译和文本规划等。

### 5.2.1 自动摘要

Hirao 等人<sup>[10]</sup>尝试使用裁剪篇章结构树进行自动摘要生成。它先使用篇章结构分析器分析出给定文章的 RST 风格的篇章结构树,然后将 RST 风格的篇章结构树转换为依存形式的篇章结构树,最后通过裁剪算法获得最终的摘要。而后, Yoshida 等人<sup>[11]</sup>指出这种两步走的方法会增加级联错误,于是通过直接学习依存形式的篇章结构树,从而只需一步即可获得文章摘要。Isonuma 等人<sup>[12]</sup>将隐式的篇章结构信息应用到自动摘要中。Xu 等人<sup>[14]</sup>将篇章结构树信息与图神经网络相结合,在抽取式自动摘要上取得了良好的性能。Wei 等人<sup>[15]</sup>将篇章关系通过图卷积网络(graph convolutional network, GCN)融入到生成式自动摘要任务中,也在生成式自动摘要上取得了良好的性能。以上研究工作均为单文档的摘要,也有一些工作利用篇章结构树进行多文档摘要。Zahri 等人<sup>[16]</sup>利用两句之间存在的修辞关系将相似的句子分成多个聚类,以识别包含共同信息的主题并从中提取出候选摘要句。然后,该方法利用条件马尔可夫随机游走模型计算每个候选摘要句的显著性,从而挑选出目标摘要句。付颖等人<sup>[17]</sup>利用宏观篇章结构信息,构建了章节和章节之间的篇章联系,改进了科技论文摘要的性能表现。在汉语中,张迎等人<sup>[13]</sup>利用篇章结构中的主次关系识别与文本摘要生成联合学习,在汉语新闻摘要任务中取得了良好的性能。

### 5.2.2 机器翻译

Tu 等人<sup>[18]</sup>提出包含 3 个步骤的翻译框架,该框架利用篇章关系辅助机器翻译任务。第 1 步,每个源语句首先被解析成一个篇章结构树。第 2 步,从篇章结构树中定义并提取翻译规则,通过双语单词对齐生成目标字符串。第 3 步,使用提取的翻译规则进行翻译从而得到最终的翻译句子。Joty 等人<sup>[19]</sup>证明句子级篇章结构子树的信息可以帮助改进现有的机器翻译评价指标,但是文档级别的篇章结构树的应用仍然是一个瓶颈。亢晓勉等人<sup>[20]</sup>则通过融入篇章结构的位置编码进一步改进了段落级的机器翻译的性能。

### 5.2.3 文本规划

篇章结构也被用于进行文本规划,如自动写作、自动问答等任务。Marcu<sup>[150]</sup>基于修辞结构理论第一次尝试进行文本规划任务。他提出一个数据驱动、自底向上的文本规划器,其输入包含篇章单元序列和每对篇章单元之间的修辞关系。虽然该方法通过制定的规则进行文本规划,而且只包含 5 种篇章关系,但是为后来利用修辞结构理论进行文本规划奠定了基础。Forsbom<sup>[151]</sup>提出包含顶层规划与微观规划的完整的文本规划器。Cojocar 等人<sup>[152]</sup>使用修辞结构理论帮助模型从本体到自然语言的生成。Galitsky 等人<sup>[153]</sup>利用篇章结构树辅助对话机器人的问答生成,用于更快的寻找到目标的答案。他们通过构建段落的篇章结构树,并利用树核(tree kernel)学习问题和答案的篇章树从而获得连贯的答案。在汉语中,刘茂福等人<sup>[154]</sup>在使用卷积神经网络和基于规则抽取关键句子后,利用篇章结构生成了 ROUGE 值更高的足球新闻报道。

## 6 英汉篇章结构分析研究异同

通过对于英语篇章结构分析理论、语料库、方法和应用的分析,可以发现英汉篇章结构分析研究有以下相同之处。

在篇章结构分析理论中,汉语研究主要借鉴英语的篇章结构分析理论相关研究,并结合汉语特点构建汉语篇章结构分析理论。其构建的语料库形式也与英语的语料库形式一致,主要为层次化篇章关系语料库和扁平化篇章关系语料库。

在篇章结构分析方法研究中,汉语逐渐与英语同步。例如在篇章结构树构建中,构建篇章结构树方法主要为移



进规约算法和自上而下算法. 并且, 英语和汉语的研究重点都在以下 3 个方面: (1) 侧重于增强篇章单元语义表示; (2) 增强篇章单元间的语义交互; (3) 更优的篇章结构树构建方法.

然而, 由于汉语和英语的语言特性不同, 它们之间也存在以下不同.

在篇章结构分析理论中, 尽管英语和汉语对于篇章关系的定义上有重合之处, 但是各个理论还是根据各自语言的特点进行了额外的关系定义, 因此其关系数量和划分并不一致, 如表 2 所示. 另外, 在英语中存在图型的篇章结构表示理论如 SDRT 和 CST, 而在汉语中还并没有相应的理论被构建. 这可能是由于图型的篇章结构的可计算性更弱<sup>[45]</sup>造成的.

在篇章结构分析方法研究中, 英语一般以子句作为基本篇章单元构建整篇文章的层次化的篇章结构, 而汉语目前还是在段落内部和段落之间分开构建篇章结构, 并没有构建从子句到整篇文章的完整篇章结构. 相比较英语的研究重点转向多任务跨语言、半监督无监督方向, 汉语篇章结构分析研究还缺少相应的尝试. 在对于篇章关系的语义理解上, 英语研究趋势从理解两个篇章单元的语义向理解篇章关系本身的语义等更深层的理解语义方向发展, 而汉语还停留在增强语义篇章单元表示和增强篇章单元语义交互上, 缺少对于篇章关系间接建模和直接建模. 在篇章结构分析应用上, 英语研究较为广泛. 但是汉语中利用篇章结构分析的应用目前处于尝试阶段, 仍然需要进一步研究如何更好地辅助下游任务.

## 7 趋势与挑战

### 7.1 趋势

在英语中, 篇章结构分析研究已经较为成熟. 各种篇章理论从不同的角度表示篇章结构. 其差异主要在篇章结构的表现形式 (浅层、树型和图型) 和篇章关系的定义上. 在语料库构建任务中, 目前的研究重心逐渐倾向于构建大规模、口语化和多语种的篇章语料库. 篇章结构树构建方法逐渐从传统的有监督机器学习和神经网络方法向多任务或跨语言、半监督或无监督等方法转换. 具体而言, 主要研究重心转移至更多样化的训练语料、增强篇章单元语义表示、增强篇章单元间的语义交互、更优的篇章结构树构建方法和更多样化的学习目标 5 个方面. 目前, 在标准基本篇章单元分割的情况下, 针对一篇文章的篇章结构分析结果的性能已经接近于人类水平.

在篇章关系识别任务中, 最近的隐式篇章关系识别的工作倾向于通过图神经网络或者多任务学习的方式获得更好的语义交互. 而且其研究重心逐渐从理解篇章单元语义向理解篇章关系语义等更深层的语义理解转移. 尽管隐式篇章关系识别近年来取得了一系列的成就, 但是其识别性能仍然有较大的提升空间.

在应用到下游任务中, 研究趋势有两个方面: (1) 逐渐从添加篇章结构信息和篇章关系信息到将篇章结构分析任务与目标任务联合学习再到将篇章结构融入到模型的设计之中过渡; (2) 逐渐从辅助文本理解向辅助文本生成任务过渡. 而且, Ko 等人<sup>[155]</sup>指出目前流行的预训练模型还不能捕获篇章关系. Koto 等人<sup>[156]</sup>也利用多个预训练模型在各种篇章任务中进行实验, 以检验预训练模型是否已经包含了篇章结构和关系等信息, 但是没有对于篇章结构树构建任务进行探究.

在汉语中, 篇章结构分析还处于发展阶段. 汉语篇章结构分析理论一般是结合英语等主流语言学理论与汉语特点而构建. 在语料库构建工作中, 目前所构建的语料库的规模都较小, 缺乏从子句到文档级别的完整篇章语料库, 而且其表现形式不够多样和完善. 篇章结构分析工作主要从篇章结构树构建和篇章关系识别两个方面进行研究. 在微观层面的研究较多, 拥有了完整的篇章结构分析器; 而对于宏观篇章结构层面, 刚刚构建完成完整的端到端系统, 对于宏观篇章结构分析还需要进一步深入研究. 在应用阶段, 无论是微观还是宏观层面都缺少相关的应用实践.

另外, 随着多语言预训练模型的提出, 打破了不同语言间的语义表示间隔. 一些研究侧重于两种语言或者多语言统一的篇章结构分析. 因此, 将英语与汉语篇章结构分析相结合也是目前研究的趋势之一.

### 7.2 挑战

综上所述, 本文认为在汉语篇章结构分析工作中对于篇章分析理论、篇章语料库建设、篇章结构分析研究和篇章结构应用等 4 个方面存在如下挑战.



● 篇章分析理论. 相较于英语篇章分析理论的多元化发展, 汉语的篇章分析理论还不够成熟和完善, 多数还处于将其他语言的篇章理论迁移到汉语中, 尚未形成针对汉语篇章从句子级别到篇章级别的统一框架, 因此对于汉语篇章结构的解释还不够透彻和深入, 需要针对当前已有的篇章结构分析理论进一步研究分析, 从中发现共性与不足.

● 篇章语料库建设. 英语篇章语料库拥有从子句到文档的完整篇章语料库, 也拥有通过情感分析远程监督方式标注的大规模篇章语料库. 而目前汉语篇章语料库的标注规模都还较小, 体裁单一. 尤其是在目前有监督的深度学习方法需要依赖大量的标注语料的情况下, 有必要扩充篇章语料库的标注规模和多样性. 因此存在以下两个方面需要进一步完善.

(1) 完整性. 一方面需要统一微观层面和宏观层面的汉语篇章语料库, 形成完整的从子句到文档级别的篇章语料库. 另一方面, 现有的语料库规模还比较小, 需要进一步扩大标注规模. 然而, 相比较如句法分析、命名实体识别等颗粒度较细的任务, 篇章结构分析语料库的基本篇章单元颗粒度较大, 标注成本较高, 较难进行大规模的人工语料标注. 因此, 需要通过设计合适的众包方法简化标注任务, 采用远程监督或者半监督的方式构建大规模篇章语料库.

(2) 多样性. 现有的汉语篇章语料库的体裁较为单一, 大多是新闻体裁, 缺少跨领域、跨体裁的篇章语料库. 因此, 在标注体裁的选取上可以从新闻体裁向小说、网络媒体、评论、法律等体裁扩展, 以此验证篇章结构分析理论的通用性. 也可以比较不同体裁下的篇章结构的差异, 为以后的篇章结构分析研究打下基础, 从而更好地辅助下游任务, 如法律文书摘要等. 另外, 多模态的信息对于理解篇章语义有很大帮助. 在多模态的篇章结构分析上的研究还较少, 例如 Akula 等人<sup>[157]</sup>构建了一个包含 310 个视频片段的视频篇章语料库. 而在汉语篇章分析研究中还处于空白阶段, 因此可以尝试进行多模态篇章结构分析的相关工作.

● 篇章结构分析研究. 在篇章结构分析研究方法中, 还应该考虑以下 5 种融合方向.

(1) 融合全局语义信息与局部语义信息. 在英语中, 构建篇章结构树的方法共 4 种: 自下而上, 基于 CYK, 基于转移和自上而下. 而汉语的主流工作则使用基于转移和自上而下两种方法. 不同的篇章结构树构建方法各有优劣, 因此需要寻找一个既能考虑全局信息, 又能增强局部匹配, 而且能够快速实用的篇章结构分析器就非常重要.

(2) 融合建模篇章单元和篇章关系表示. 对于篇章关系识别任务, 当前研究重点逐渐从建模篇章单元表示, 到建模篇章单元间的语义交互, 再到建模篇章关系的表示, 因此整体朝着深入理解篇章关系语义发展. 在接下来的研究中, 应该倾向于直接建模篇章关系, 如利用对于篇章关系的解释或者联合建模篇章单元的语义表示和篇章关系的语义表示辅助篇章关系识别.

(3) 融合篇章结构树构建、主次识别、关系识别等任务. 当前的工作都是将篇章结构树构建、主次识别和关系识别等任务当作管道模型, 没有将各个任务联合起来. 在 RST-DT 中, 更多的工作注重篇章结构树的构建, 对于主次识别和关系识别没有足够重视, 一般使用多层感知机直接进行分类. 在 PDTB 中, 更多的工作则侧重于篇章关系识别, 没有对层次化结构进行考虑. 然而, 篇章结构树的构建、主次识别和关系识别任务也能够相互补充, 在早期的工作中也有对于篇章结构树构建和主次识别, 主次和关系识别联合建模的尝试, 未来工作应该深入挖掘篇章结构树构建、主次识别和关系识别任务间的联系.

(4) 融合微观和宏观篇章结构. 在宏观篇章结构分析中, 基本篇章单元为段落. 一般的, 段落长度比句子更长, 更好的表示段落内以及多个段落中重要的信息是一个挑战. 而且, 恰当的融合微观和宏观篇章结构也能够对于篇章单元表示有所帮助. 另外, 段落内部包含信息冗余, 段落之间没有显式连接词, 段落与段落之间的联系不如句子与句子之间联系的密切, 这会造成篇章关系比较模糊, 从而难以识别宏观篇章关系. 因此需要更好的构建两个段落之间的语义交互, 例如通过融合宏观的全局信息与微观的局部信息增强篇章单元间的语义交互.

(5) 融合多语言或多任务篇章结构分析. 针对篇章语料库语料较为稀少的问题, 除了扩充语料库规模外, 也可以通过与其他语言的篇章语料库语料联合训练, 或者利用多语言预训练模型与机器翻译方法统一篇章单元语义表示进行篇章结构分析. 另外, 借鉴半监督或无监督的句法解析方法进行篇章结构分析也是目前研究需要探索的方向之一.

● 篇章结构应用. 相比较英语中篇章结构信息在下游任务中的广泛应用, 汉语目前还处于尝试阶段. 一方面由

于篇章结构分析的结果在汉语中还不够良好, 受限于语料库的体裁和规模, 较难应用到其他任务中. 另一方面尚未找到合适的将篇章结构信息应用到下游任务中的方法. 因此, 在此方面应该进行更多的尝试, 包括利用篇章结构更好表示文本语义等方向, 如在自动摘要、作文打分、新闻生成等任务上进行深入探究.

## 8 总结与展望

本文梳理总结了英汉篇章结构分析研究的相关工作, 特别是针对英语中的层次化篇章关系语料库 (RST-DT) 和扁平化篇章关系语料库 (PDTB) 以及汉语中的层次化篇章关系语料库 (SUDA-CDTB 和 MCDTB) 和扁平化篇章关系语料库 (HIT-CDTB 和 LDC-CDTB) 上的相关工作进行了重点介绍. 本文将 RST-DT 相关工作的研究重点归纳为更多样化的训练语料、增强篇章单元的表达、增强篇章单元间的语义交互、更优的篇章结构树构建方法与更多样化的学习目标等 5 个部分. 而且, 本文将在 PDTB 相关工作的研究重点归纳为增强篇章单元的表达、增强篇章单元间的语义交互以及联合建模篇章关系与其他语义任务 3 个部分. 接着, 本文归纳总结了目前汉语篇章结构分析上的相关进展, 特别是层次化篇章关系语料库 (SUDA-CDTB 和 MCDTB) 和扁平化篇章关系语料库 (HIT-CDTB 和 LDC-CDTB) 的相关工作, 并指出目前汉语篇章结构分析虽然取得一定成果, 但是在篇章分析理论、篇章语料库、篇章结构分析研究、篇章结构应用等方面仍然存在诸多挑战, 为接下来的篇章结构分析工作提供指导和帮助.

### References:

- [1] Mann WC, Thompson SA. *Rhetorical Structure Theory: A Theory of Text Organization*. Los Angeles: University of Southern California, Information Sciences Institute, 1987.
- [2] Li YC. *Research of Chinese discourse structure representation and resource construction* [Ph.D. Thesis]. Suzhou: Soochow Univeristy, 2015 (in Chinese with English abstract).
- [3] Li YC, Feng WH, Sun J, Kong F, Zhou GD. Building Chinese discourse corpus with connective-driven dependency tree structure. In: *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Doha: ACL, 2014. 2105–2114. [doi: [10.3115/v1/D14-1224](https://doi.org/10.3115/v1/D14-1224)]
- [4] Radev DR. A common theory of information fusion from multiple text sources step one: Cross-document structure. In: *Proc. of the 1st SIGdial Workshop on Discourse and Dialogue*. Hong Kong: ACL, 2000. 74–83. [doi: [10.3115/1117736.1117745](https://doi.org/10.3115/1117736.1117745)]
- [5] Asher N, Lascarides A. *Logics of Conversation*. Cambridge: Cambridge University Press, 2003.
- [6] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 2011, 37(2): 267–307. [doi: [10.1162/COLI\\_a\\_00049](https://doi.org/10.1162/COLI_a_00049)]
- [7] Heerschop B, Goossen F, Hogenboom A, Frasinca F, Kaymak U, de Jong F. Polarity analysis of texts using discourse structure. In: *Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management (CIKM)*. Glasgow: ACM, 2011. 1061–1070. [doi: [10.1145/2063576.2063730](https://doi.org/10.1145/2063576.2063730)]
- [8] Fu XH, Liu WW, Xu YY, Yu C, Wang T. Long short-term memory network over rhetorical structure theory for sentence-level sentiment analysis. In: *Proc. of the 8th Asian Conf. on Machine Learning (ACML)*. Hamilton: JMLR.org, 2016. 17–32.
- [9] Kraus M, Feuerriegel S. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 2019, 118: 65–79. [doi: [10.1016/j.eswa.2018.10.002](https://doi.org/10.1016/j.eswa.2018.10.002)]
- [10] Hirao T, Yoshida Y, Nishino M, Yasuda N, Nagata M. Single-document summarization as a tree knapsack problem. In: *Proc. of the 2013 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle: Association for Computational Linguistics, 2013. 1515–1520.
- [11] Yoshida Y, Suzuki J, Hirao T, Nagata M. Dependency-based discourse parser for single-document summarization. In: *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Doha: Association for Computational Linguistics, 2014. 1834–1839. [doi: [10.3115/v1/D14-1196](https://doi.org/10.3115/v1/D14-1196)]
- [12] Isonuma M, Mori J, Sakata I. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Florence: Association for Computational Linguistics, 2019. 2142–2152. [doi: [10.18653/v1/P19-1206](https://doi.org/10.18653/v1/P19-1206)]
- [13] Zhang Y, Wang ZQ, Wang HL. Single document extractive summarization with satellite and nuclear relations. *Journal of Chinese Information Processing*, 2019, 33(8): 67–76 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2019.08.009](https://doi.org/10.3969/j.issn.1003-0077.2019.08.009)]

- [14] Xu JC, Gan Z, Cheng Y, Liu JJ. Discourse-aware neural extractive text summarization. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Association for Computational Linguistics, 2020. 5021–5031. [doi: [10.18653/v1/2020.acl-main.451](https://doi.org/10.18653/v1/2020.acl-main.451)]
- [15] Wei WJ, Wang HL, Wang ZQ. Abstractive summarization via discourse relation and graph convolutional networks. In: Proc. of the 9th Natural Language Processing and Chinese Computing (NLPCC). Zhengzhou: Springer, 2020. 331–342. [doi: [10.1007/978-3-030-60457-8\\_27](https://doi.org/10.1007/978-3-030-60457-8_27)]
- [16] Zahri NAH, Fukumoto F, Suguru M, Lynn OB. Exploiting rhetorical relations to multiple documents text summarization. *Int'l Journal of Network Security & Its Applications*, 2015, 7(2): 1–22. [doi: [10.5121/ijnsa.2015.7201](https://doi.org/10.5121/ijnsa.2015.7201)]
- [17] Fu Y, Wang HL, Wang ZQ. Scientific paper summarization model using macro discourse structure. *Journal of Computer Applications*, 2021, 41(10): 2864–2870 (in Chinese with English abstract). [doi: [10.11772/j.issn.1001-9081.2020121945](https://doi.org/10.11772/j.issn.1001-9081.2020121945)]
- [18] Tu M, Zhou Y, Zong CQ. A novel translation framework based on rhetorical structure theory. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia: Association for Computational Linguistics, 2013. 370–374.
- [19] Joty S, Guzmán F, Márquez L, Nakov P. Discourse structure in machine translation evaluation. *Computational Linguistics*, 2017, 43(4): 683–722. [doi: [10.1162/COLI\\_a\\_00298](https://doi.org/10.1162/COLI_a_00298)]
- [20] Kang XM, Zong CQ. Fusion of discourse structural position encoding for neural machine translation. *Chinese Journal of Intelligent Science and Technology*, 2020, 2(2): 144–152 (in Chinese with English abstract). [doi: [10.11959/j.issn.2096-6652.202016](https://doi.org/10.11959/j.issn.2096-6652.202016)]
- [21] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). New Orleans: ACL, 2018. 2227–2237. [doi: [10.18653/v1/N18-1202](https://doi.org/10.18653/v1/N18-1202)]
- [22] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Minneapolis: ACL, 2019. 4171–4186. [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
- [23] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019, 1(8): 9–33.
- [24] Yang ZL, Dai ZH, Yang YM, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 517.
- [25] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou YQ, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020, 21(140): 1–67.
- [26] Zhang ZY, Han X, Liu ZY, Jiang X, Sun MS, Liu Q. ERNIE: Enhanced language representation with informative entities. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence: ACL, 2019. 1441–1451. [doi: [10.18653/v1/P19-1139](https://doi.org/10.18653/v1/P19-1139)]
- [27] Liu WJ, Zhou P, Zhao Z, Wang ZR, Ju Q, Deng HT, Wang P. K-BERT: Enabling language representation with knowledge graph. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence (AAAI). New York: AAAI, 2020. 2901–2908. [doi: [10.1609/aaai.v34i03.5681](https://doi.org/10.1609/aaai.v34i03.5681)]
- [28] Zhang MS, Li ZH, Fu GH, Zhang M. Dependency-based syntax-aware word representations. *Artificial Intelligence*, 2021, 292: 103427. [doi: [10.1016/j.artint.2020.103427](https://doi.org/10.1016/j.artint.2020.103427)]
- [29] Carlson L, Marcu D, Okurowski ME. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt J, Smith RW, eds. *Current and New Directions in Discourse and Dialogue*. Dordrecht: Springer, 2003. 85–112. [doi: [10.1007/978-94-010-0019-2\\_5](https://doi.org/10.1007/978-94-010-0019-2_5)]
- [30] Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi A, Webber B. The penn discourse TreeBank 2.0. In: Proc. of the 6th Int'l Conf. on Language Resources and Evaluation (LREC). Marrakech: European Language Resources Association, 2008. 2961–2968.
- [31] Jiang F, Xu S, Chu XM, Li PF, Zhu QM, Zhou GD. MCDTB: A macro-level Chinese discourse TreeBank. In: Proc. of the 27th Int'l Conf. on Computational Linguistics (COLING). Santa Fe: ACL, 2018. 3493–3504.
- [32] Jiang F. Research on the Method of Chinese macro discourse resources construction and structure analysis [MS. Thesis]. Suzhou: Soochow Univeristy, 2019 (in Chinese with English abstract). [doi: [10.27351/d.cnki.gszzhu.2019.000756](https://doi.org/10.27351/d.cnki.gszzhu.2019.000756)]
- [33] Zhang MY, Qin B, Liu T. Chinese discourse relation semantic taxonomy and annotation. *Journal of Chinese Information Processing*, 2014, 28(2): 28–36 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2014.02.004](https://doi.org/10.3969/j.issn.1003-0077.2014.02.004)]
- [34] Zhou YP, Xue NW. The Chinese Discourse TreeBank: A Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 2015, 49(2): 397–431. [doi: [10.1007/s10579-014-9290-3](https://doi.org/10.1007/s10579-014-9290-3)]
- [35] Stede M, Neumann A. Potsdam commentary corpus 2.0: Annotation for discourse research. In: Proc. of the 9th Int'l Conf. on Language

- Resources and Evaluation (LREC). Reykjavik: European Language Resources Association, 2014. 925–929.
- [36] Redeker G, Berzlánovich I, Van Der Vliet N, Bouma G, Egg M. Multi-layer discourse annotation of a Dutch text corpus. In: Proc. of the 8th Int'l Conf. on Language Resources and Evaluation (LREC). Istanbul: European Language Resources Association, 2012. 2820–2825.
- [37] Iruškieta M, Braud C. EusDisParser: Improving an under-resourced discourse parser with cross-lingual data. In: Proc. of the 2019 Workshop on Discourse Relation Parsing and Treebanking. Minneapolis: ACL, 2019. 62–71. [doi: [10.18653/v1/W19-2709](https://doi.org/10.18653/v1/W19-2709)]
- [38] Da Cunha I, Torres-Moreno JM, Sierra G. On the development of the RST spanish Treebank. In: Proc. of the 5th Linguistic Annotation Workshop. Portland: ACL, 2011. 1–10.
- [39] Hou SL, Zhang SH, Fei CQ. Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, 2020, 157: 113421. [doi: [10.1016/j.eswa.2020.113421](https://doi.org/10.1016/j.eswa.2020.113421)]
- [40] Yan WR, Xu Y, Zhu SS, Hong Y, Yao JM, Zhu QM. A survey to discourse relation analyzing. *Journal of Chinese Information Processing*, 2016, 30(4): 1–11 (in Chinese with English abstract).
- [41] Morey M, Muller P, Asher N. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 2018, 44(2): 197–235. [doi: [10.1162/COLI\\_a\\_00314](https://doi.org/10.1162/COLI_a_00314)]
- [42] Song W, Liu LZ. Representation learning in discourse parsing: A survey. *Science China Technological Sciences*, 2020, 63(10): 1921–1946. [doi: [10.1007/s11431-020-1685-2](https://doi.org/10.1007/s11431-020-1685-2)]
- [43] Li JQ, Liu M, Qin B, Liu T. A survey of discourse parsing. *Frontiers of Computer Science*, 2022, 16(5): 165329. [doi: [10.1007/s11704-021-0500-z](https://doi.org/10.1007/s11704-021-0500-z)]
- [44] Xu F, Zhu QM, Zhou GD. Survey of discourse analysis methods. *Journal of Chinese Information Processing*, 2013, 27(3): 20–32, 55. (in Chinese with English abstract).
- [45] Kang XM, Zong CQ, Xue NW. A survey of discourse representations for Chinese discourse annotation. *ACM Trans. on Asian and Low-resource Language Information Processing*, 2019, 18(3): 26. [doi: [10.1145/3293442](https://doi.org/10.1145/3293442)]
- [46] Kong F, Wang HL, Zhou GD. Survey on Chinese discourse understanding. *Ruan Jian Xue Bao/Journal of Software*, 2019, 30(7): 2052–2072 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/5834.htm> [doi: [10.13328/j.cnki.jos.005834](https://doi.org/10.13328/j.cnki.jos.005834)]
- [47] Hobbs JR. Why is discourse coherent. Technical Report, SRI Int'l Menlo Park CA. 1978.
- [48] Webber B. D-LTAG: Extending lexicalized tag to discourse. *Cognitive Science*, 2004, 28(5): 751–779. [doi: [10.1207/s15516709cog2805\\_6](https://doi.org/10.1207/s15516709cog2805_6)]
- [49] Lascarides A, Asher N. Segmented discourse representation theory: Dynamic semantics with discourse structure. In: Bunt H, Muskens R, eds. *Computing Meaning*. Dordrecht: Springer, 2008. 87–124. [doi: [10.1007/978-1-4020-5958-2\\_5](https://doi.org/10.1007/978-1-4020-5958-2_5)]
- [50] Wu WZ, Tian XL. *Chinese Sentence Group*. Beijing: The Commercial Press, 2000 (in Chinese).
- [51] Xing FY. *Research on Chinese Complex Sentences*. Beijing: The Commercial Press, 2001 (in Chinese).
- [52] Hao XY, Liu W, Li R, Liu KY. Description systems of the Chinese FrameNet database and software tools. *Journal of Chinese Information Processing*, 2007, 21(5): 96–100, 138 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2007.05.018](https://doi.org/10.3969/j.issn.1003-0077.2007.05.018)]
- [53] Chu XM. *Research on representation schema, resource construction and computational modeling of macro discourse structure* [Ph.D. Thesis]. Suzhou: Soochow Univeristy, 2018 (in Chinese with English abstract). [doi: [10.27351/d.cnki.gszzhu.2018.000122](https://doi.org/10.27351/d.cnki.gszzhu.2018.000122)]
- [54] Chu XM, Xi XF, Jiang F, Xu S, Zhu QM, Zhou GD. Macro discourse structure representation schema and corpus construction. *Ruan Jian Xue Bao/Journal of Software*, 2020, 31(2): 321–343 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/5868.htm> [doi: [10.13328/j.cnki.jos.005868](https://doi.org/10.13328/j.cnki.jos.005868)]
- [55] Joshi AK, Schabes Y. *Tree-adjointing grammars and lexicalized grammars*. Philadelphia: University of Pennsylvania, 1991. 445.
- [56] van Dijk TA. *News as Discourse*. Hillsdale: Lawrence Erlbaum Associates, 1988.
- [57] Subba R, Di Eugenio B. An effective discourse parser that uses rich linguistic information. In: Proc. of the 2009 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). Boulder: ACL, 2009. 566–574.
- [58] Zeldes A. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 2017, 51(3): 581–612. [doi: [10.1007/s10579-016-9343-x](https://doi.org/10.1007/s10579-016-9343-x)]
- [59] Huber P, Carenini G. MEGA RST discourse Treebanks with structure and nuclearity from scalable distant sentiment supervision. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 7442–7457. [doi: [10.18653/v1/2020.emnlp-main.603](https://doi.org/10.18653/v1/2020.emnlp-main.603)]
- [60] Asher N, Hunter J, Morey M, Farah B, Afantenos S. Discourse structure and dialogue acts in multiparty dialogue: The STAC corpus. In: Proc. of the 10th Int'l Conf. on Language Resources and Evaluation (LREC). Portorož: European Language Resources Association, 2016. 2721–2727.

- [61] Li JQ, Liu M, Kan MY, Zheng ZH, Wang ZK, Lei WQ, Liu T, Qin B. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In: Proc. of the 28th Int'l Conf. on Computational Linguistics (COLING). Barcelona: International Committee on Computational Linguistics, 2020. 2642–2652. [doi: [10.18653/v1/2020.coling-main.238](https://doi.org/10.18653/v1/2020.coling-main.238)]
- [62] Ma MD, Bowden KK, Wu JQ, Cui W, Walker M. Implicit discourse relation identification for open-domain dialogues. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence: ACL, 2019. 666–672. [doi: [10.18653/v1/P19-1065](https://doi.org/10.18653/v1/P19-1065)]
- [63] Zeyrek D, Mendes A, Grishina Y, Kurfalı M, Gibbon S, Ogrodniczuk M. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources and Evaluation*, 2020, 54(2): 587–613. [doi: [10.1007/s10579-019-09445-9](https://doi.org/10.1007/s10579-019-09445-9)]
- [64] Li J, Sun AX, Joty S. SEGBOT: A generic neural text segmentation model with pointer network. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence (IJCAI). Stockholm: AAAI, 2018. 4166–4172.
- [65] Lin X, Joty S, Jwalapuram P, Bari MS. A unified linear-time framework for sentence-level discourse parsing. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence: ACL, 2019. 4190–4200. [doi: [10.18653/v1/P19-1410](https://doi.org/10.18653/v1/P19-1410)]
- [66] Nguyen TT, Nguyen XP, Joty S, Li XL. RST Parsing from scratch. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). ACL, 2021. 1613–1625. [doi: [10.18653/v1/2021.naacl-main.128](https://doi.org/10.18653/v1/2021.naacl-main.128)]
- [67] Hernault H, Prendinger H, du Verle DA, Ishizuka M. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 2010, 1(3): 1–33. [doi: [10.5087/dad.2010.003](https://doi.org/10.5087/dad.2010.003)]
- [68] Feng VW, Hirst G. A linear-time bottom-up discourse parser with constraints and post-editing. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore: ACL, 2014. 511–521. [doi: [10.3115/v1/P14-1048](https://doi.org/10.3115/v1/P14-1048)]
- [69] Joty S, Carenini G, Ng R, Mehdad Y. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In: Proc. of the 51st Annual Meeting of the Association for Computational Linguistics (ACL). Sofia: ACL, 2013. 486–496.
- [70] Joty S, Carenini G, Ng R. A novel discriminative framework for sentence-level discourse analysis. In: Proc. of the 2012 Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP). Jeju Island: ACL, 2012. 904–915.
- [71] Ji YF, Eisenstein J. Representation learning for text-level discourse parsing. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL). Baltimore: ACL, 2014. 13–24. [doi: [10.3115/v1/P14-1002](https://doi.org/10.3115/v1/P14-1002)]
- [72] Heilman M, Sagae K. Fast rhetorical structure theory discourse parsing. arXiv:1505.02425, 2015.
- [73] Wang YZ, Li SJ, Wang HF. A two-stage parsing method for text-level discourse analysis. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver: ACL, 2017. 184–188. [doi: [10.18653/v1/P17-2029](https://doi.org/10.18653/v1/P17-2029)]
- [74] Li JW, Li RM, Hovy E. Recursive deep models for discourse parsing. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Doha: ACL, 2014. 2061–2069. [doi: [10.3115/v1/D14-1220](https://doi.org/10.3115/v1/D14-1220)]
- [75] Li Q, Li TS, Chang BB. Discourse parsing with attention-based hierarchical neural networks. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Austin: ACL, 2016. 362–371. [doi: [10.18653/v1/D16-1035](https://doi.org/10.18653/v1/D16-1035)]
- [76] Jia YY, Ye Y, Feng YS, Lai YX, Yan R, Zhao DY. Modeling discourse cohesion for discourse parsing via memory network. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). Melbourne: ACL, 2018. 438–443. [doi: [10.18653/v1/P18-2070](https://doi.org/10.18653/v1/P18-2070)]
- [77] Yu N, Zhang MS, Fu GH. Transition-based neural RST parsing with implicit syntax features. In: Proc. of the 27th Int'l Conf. on Computational Linguistics (COLING). Santa Fe: ACL, 2018. 559–570.
- [78] Mabona A, Rimell L, Clark S, Vlachos A. Neural generative rhetorical structure parsing. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 2284–2295. [doi: [10.18653/v1/D19-1233](https://doi.org/10.18653/v1/D19-1233)]
- [79] Zhang LY, Tan X, Kong F, Zhou GD. A recursive information flow gated model for RST-style text-level discourse parsing. In: Proc. of the 8th CCF Int'l Conf. on Natural Language Processing and Chinese Computing (NLPCC). Dunhuang: Springer, 2019. 231–241. [doi: [10.1007/978-3-030-32236-6\\_20](https://doi.org/10.1007/978-3-030-32236-6_20)]
- [80] Li JF, Xiao L. Tree representations in transition system for RST parsing. In: Proc. of the 28th Int'l Conf. on Computational Linguistics (COLING). Barcelona: International Committee on Computational Linguistics, 2020. 6746–6751. [doi: [10.18653/v1/2020.coling-main.593](https://doi.org/10.18653/v1/2020.coling-main.593)]
- [81] Liu LL, Lin X, Joty S, Han SM, Bing LD. Hierarchical pointer net parsing. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019.



- 1006–1016. [doi: [10.18653/v1/D19-1093](https://doi.org/10.18653/v1/D19-1093)]
- [82] Kobayashi N, Hirao T, Kamigaito H, Okumura M, Nagata M. Top-down RST parsing utilizing granularity levels in documents. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence (AAAI). New York: AAAI, 2020. 8099–8106. [doi: [10.1609/aaai.v34i05.6321](https://doi.org/10.1609/aaai.v34i05.6321)]
- [83] Koto F, Lau JH, Baldwin T. Top-down discourse parsing via sequence labelling. In: Proc. of the 16th Conf. of the European Chapter of the Association for Computational Linguistics (EACL). ACL, 2021. 715–726. [doi: [10.18653/v1/2021.eacl-main.60](https://doi.org/10.18653/v1/2021.eacl-main.60)]
- [84] Braud C, Plank B, Søgaard A. Multi-view and multi-task training of RST discourse parsers. In: Proc. of the 26th Int'l Conf. on Computational Linguistics (COLING). Osaka: The COLING 2016 Organizing Committee, 2016. 1903–1913.
- [85] Braud C, Coavoux M, Søgaard A. Cross-lingual RST discourse parsing. In: Proc. of the 15th Conf. of the European Chapter of the Association for Computational Linguistics. Valencia: ACL, 2017. 292–304.
- [86] Coavoux M, Crabbé B. Neural greedy constituent parsing with dynamic oracles. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (ACL). Berlin: ACL, 2016. 172–182. [doi: [10.18653/v1/P16-1017](https://doi.org/10.18653/v1/P16-1017)]
- [87] Liu ZY, Shi K, Chen N. Multilingual neural RST discourse parsing. In: Proc. of the 28th Int'l Conf. on Computational Linguistics (COLING). Barcelona: Int'l Committee on Computational Linguistics, 2020. 6730–6738. [doi: [10.18653/v1/2020.coling-main.591](https://doi.org/10.18653/v1/2020.coling-main.591)]
- [88] Guz G, Carenini G. Coreference for discourse parsing: A neural approach. In: Proc. of the 1st Workshop on Computational Approaches to Discourse. ACL, 2020. 160–167. [doi: [10.18653/v1/2020.codi-1.17](https://doi.org/10.18653/v1/2020.codi-1.17)]
- [89] Kobayashi N, Hirao T, Nakamura K, Kamigaito H, Okumura M, Nagata M. Split or merge: Which is better for unsupervised RST parsing? In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: ACL, 2019. 5801–5806. [doi: [10.18653/v1/D19-1587](https://doi.org/10.18653/v1/D19-1587)]
- [90] Nishida N, Nakayama H. Unsupervised discourse constituency parsing using Viterbi EM. *Trans. of the Association for Computational Linguistics*, 2020, 8: 215–230. [doi: [10.1162/tacl\\_a\\_00312](https://doi.org/10.1162/tacl_a_00312)]
- [91] Huber P, Carenini G. Unsupervised learning of discourse structures using a tree autoencoder. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence (AAAI). AAAI, 2021. 13107–13115.
- [92] Kobayashi N, Hirao T, Kamigaito H, Okumura M, Nagata M. Improving neural RST parsing model with silver agreement subtrees. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). ACL, 2021. 1600–1612. [doi: [10.18653/v1/2021.naacl-main.127](https://doi.org/10.18653/v1/2021.naacl-main.127)]
- [93] Xiao W, Huber P, Carenini G. Predicting discourse trees from transformer-based neural summarizers. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). ACL, 2021. 4139–4152. [doi: [10.18653/v1/2021.naacl-main.326](https://doi.org/10.18653/v1/2021.naacl-main.326)]
- [94] Liang L, Zhao Z, Webber B. Extending implicit discourse relation recognition to the PDTB-3. In: Proc. of the 1st Workshop on Computational Approaches to Discourse. ACL, 2020. 135–147. [doi: [10.18653/v1/2020.codi-1.14](https://doi.org/10.18653/v1/2020.codi-1.14)]
- [95] Pitler E, Raghupathy M, Mehta H, Nenkova A, Lee A, Joshi A. Easily identifiable discourse relations. In: Proc. of the 22nd Int'l Conf. on Computational Linguistics (COLING). Manchester: Coling 2008 Organizing Committee, 2008. 87–90.
- [96] Pitler E, Nenkova A. Using syntax to disambiguate explicit discourse connectives in text. In: Proc. of the Joint Conf. of the 47th Annual Meeting of the ACL and the 4th Int'l Joint Conf. on Natural Language Processing of the AFNLP. Suntec: ACL, 2009. 13–16.
- [97] Zhang B, Su JS, Xiong DY, Lu YJ, Duan H, Yao JF. Shallow convolutional neural network for implicit discourse relation recognition. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Lisbon: ACL, 2015. 2230–2235. [doi: [10.18653/v1/D15-1266](https://doi.org/10.18653/v1/D15-1266)]
- [98] Ji YF, Eisenstein J. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Trans. of the Association for Computational Linguistics*, 2015, 3: 329–344. [doi: [10.1162/tacl\\_a\\_00142](https://doi.org/10.1162/tacl_a_00142)]
- [99] Qin LH, Zhang ZS, Zhao H. A stacking gated neural architecture for implicit discourse relation classification. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Austin: ACL, 2016. 2263–2270. [doi: [10.18653/v1/D16-1246](https://doi.org/10.18653/v1/D16-1246)]
- [100] Bai HX, Zhao H. Deep enhanced representation for implicit discourse relation recognition. In: Proc. of the 27th Int'l Conf. on Computational Linguistics (COLING). Santa Fe: ACL, 2018. 571–583.
- [101] Liu Y, Li SJ. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Austin: ACL, 2016. 1224–1233. [doi: [10.18653/v1/D16-1130](https://doi.org/10.18653/v1/D16-1130)]
- [102] Lan M, Wang JX, Wu YB, Niu ZY, Wang HF. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In: Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen: ACL, 2017. 1299–1308. [doi: [10.18653/v1/D17-1134](https://doi.org/10.18653/v1/D17-1134)]
- [103] Guo FY, He RF, Jin D, Dang JW, Wang LB, Li XG. Implicit discourse relation recognition using neural tensor network with interactive

- attention and sparse learning. In: Proc. of the 27th Int'l Conf. on Computational Linguistics (COLING). Santa Fe: ACL, 2018. 547–558.
- [104] Ruan HB, Hong Y, Xu Y, Huang Z, Zhou GD, Zhang M. Interactively-propagative attention learning for implicit discourse relation recognition. In: Proc. of the 28th Int'l Conf. on Computational Linguistics (COLING). Barcelona: International Committee on Computational Linguistics, 2020. 3168–3178. [doi: [10.18653/v1/2020.coling-main.282](https://doi.org/10.18653/v1/2020.coling-main.282)]
- [105] Guo FY, He RF, Dang JW, Wang J. Working memory-driven neural networks with a novel knowledge enhancement paradigm for implicit discourse relation recognition. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence (AAAI). New York: AAAI, 2020. 7822–7829. [doi: [10.1609/aaai.v34i05.6287](https://doi.org/10.1609/aaai.v34i05.6287)]
- [106] Zhang YX, Meng FD, Li P, Jian P, Zhou J. Context tracking network: Graph-based context modeling for implicit discourse relation recognition. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). ACL, 2021. 1592–1599. [doi: [10.18653/v1/2021.naacl-main.126](https://doi.org/10.18653/v1/2021.naacl-main.126)]
- [107] Shi W, Demberg V. Learning to explicitate connectives with seq2seq network for implicit discourse relation classification. In: Proc. of the 13th Int'l Conf. on Computational Semantics (IWCS). Gothenburg: ACL, 2019. 188–199. [doi: [10.18653/v1/W19-0416](https://doi.org/10.18653/v1/W19-0416)]
- [108] Nguyen LT, Van Ngo L, Than K, Nguyen TH. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence: ACL, 2019. 4201–4207. [doi: [10.18653/v1/P19-1411](https://doi.org/10.18653/v1/P19-1411)]
- [109] He RF, Wang J, Guo FY, Han YG. TransS-Driven joint learning architecture for implicit discourse relation recognition. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). ACL, 2020. 139–148. [doi: [10.18653/v1/2020.acl-main.14](https://doi.org/10.18653/v1/2020.acl-main.14)]
- [110] Long WQ, Webber B, Xiong DY. TED-CDB: A large-scale Chinese discourse relation dataset on TED Talks. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 2793–2803. [doi: [10.18653/v1/2020.emnlp-main.223](https://doi.org/10.18653/v1/2020.emnlp-main.223)]
- [111] Cheng Y, Li SJ. Zero-shot Chinese discourse dependency parsing via cross-lingual mapping. In: Proc. of the 1st Workshop on Discourse Structure in Neural NLG. Tokyo: ACL, 2019. 24–29. [doi: [10.18653/v1/W19-8104](https://doi.org/10.18653/v1/W19-8104)]
- [112] Kong F, Zhou GD. A CDT-styled end-to-end Chinese discourse parser. ACM Trans. on Asian and Low-resource Language Information Processing, 2017, 16(4): 1–17. [doi: [10.1145/3099557](https://doi.org/10.1145/3099557)]
- [113] Jia YY, Feng YS, Ye Y, Lv C, Shi CD, Zhao DY. Improved discourse parsing with two-step neural transition-based model. ACM Trans. on Asian and Low-Resource Language Information Processing, 2018, 17(2): 1–21. [doi: [10.1145/3152537](https://doi.org/10.1145/3152537)]
- [114] Sun C, Kong F. A transition-based framework for Chinese discourse structure parsing. Journal of Chinese Information Processing, 2018, 32(12): 48–56 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2018.12.006](https://doi.org/10.3969/j.issn.1003-0077.2018.12.006)]
- [115] Wang TS, Li PF, Zhu QM. A multi-stage strategy for Chinese discourse tree construction. In: Proc. of the 2019 Int'l Conf. on Asian Language Processing (IALP). Shanghai: ACL, 2019. 302–307. [doi: [10.1109/IALP48816.2019.9037684](https://doi.org/10.1109/IALP48816.2019.9037684)]
- [116] Wang TS. Research on the method of Chinese micro-discourse analysis [MS. Thesis]. Suzhou: Soochow Univeristy, 2019 (in Chinese with English abstract). [doi: [10.27351/d.cnki.gszzhu.2020.001991](https://doi.org/10.27351/d.cnki.gszzhu.2020.001991)]
- [117] Zhang LY, Xing YQ, Kong F, Li PF, Zhou GD. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). ACL, 2020. 6386–6395. [doi: [10.18653/v1/2020.acl-main.569](https://doi.org/10.18653/v1/2020.acl-main.569)]
- [118] Zhang LY, Kong F, Zhou GD. Adversarial learning for discourse rhetorical structure parsing. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing (ACL-IJCNN). ACL, 2021. 3946–3957. [doi: [10.18653/v1/2021.acl-long.305](https://doi.org/10.18653/v1/2021.acl-long.305)]
- [119] Xu S, Li PF, Zhou GD, Zhu QM. Employing text matching network to recognise nuclearity in Chinese discourse. In: Proc. of the 27th Int'l Conf. on Computational Linguistics (COLING). Santa Fe: ACL, 2018. 525–535.
- [120] Xu S. Research on micro discourse nuclearity and relation recognition in Chinese [MS. Thesis]. Suzhou: Soochow Univeristy, 2019 (in Chinese with English abstract). [doi: [10.27351/d.cnki.gszzhu.2019.000679](https://doi.org/10.27351/d.cnki.gszzhu.2019.000679)]
- [121] Wang TS, Li PF, Zhu QM. GMN-based nuclearity recognition in Chinese discourse. Journal of Chinese Information Processing, 2019, 33(5): 39–46 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2019.05.005](https://doi.org/10.3969/j.issn.1003-0077.2019.05.005)]
- [122] Xu S, Wang TS, Li PF, Zhu QM. Multi-layer attention network based Chinese implicit discourse relation recognition. Journal of Chinese Information Processing, 2019, 33(8): 12–19, 35 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2019.08.002](https://doi.org/10.3969/j.issn.1003-0077.2019.08.002)]
- [123] Xu S, Li PF, Kong F, Zhu QM, Zhou GD. Topic tensor network for implicit discourse relation recognition in Chinese. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). Florence: ACL, 2019. 608–618. [doi: [10.18653/v1/P19-1058](https://doi.org/10.18653/v1/P19-1058)]

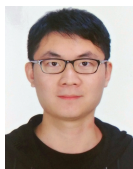
- [124] Xue NW, Zhang XH, Jiang ZX, Palmer M, Xia F, Chiou FD, Chang MY. Chinese treebank 8.0 LDC2013T21. Philadelphia: Linguistic Data Consortium, 2013.
- [125] Jiang F, Li PF, Chu XM, Zhu QM, Zhou GD. Recognizing macro Chinese discourse structure on label degeneracy combination model. In: Proc. of the 7th CCF Int'l Conf. on Natural Language Processing and Chinese Computing (NLPCC). Hohhot: Springer, 2018. 92–104. [doi: [10.1007/978-3-319-99501-4\\_8](https://doi.org/10.1007/978-3-319-99501-4_8)]
- [126] Zhou Y, Chu XM, Li PF, Zhu QM. Constructing Chinese macro discourse tree via multiple views and word pair similarity. In: Proc. of the 8th CCF Int'l Conf. on Natural Language Processing and Chinese Computing (NLPCC). Dunhuang: Springer, 2019. 773–786. [doi: [10.1007/978-3-030-32233-5\\_60](https://doi.org/10.1007/978-3-030-32233-5_60)]
- [127] Zhou Y. Research on the method of Chinese macro discourse tree auto construction [MS. Thesis]. Suzhou: Soochow Univeristy, 2020 (in Chinese with English abstract). [doi: [10.27351/d.cnki.gszhu.2020.001566](https://doi.org/10.27351/d.cnki.gszhu.2020.001566)]
- [128] Fan YX, Jiang F, Chu XM, Li PF, Zhu QM. Combining global and local information to recognize Chinese macro discourse structure. In: Proc. of the 19th Chinese National Conf. on Computational Linguistics. Haikou: Chinese Information Processing Society of China, 2020. 183–194.
- [129] Jiang F, Chu XM, Li PF, Kong F, Zhu QM. Chinese paragraph-level discourse parsing with global backward and local reverse reading. In: Proc. of the 28th Int'l Conf. on Computational Linguistics (COLING). Barcelona: Int'l Committee on Computational Linguistics, 2020. 5749–5759. [doi: [10.18653/v1/2020.coling-main.506](https://doi.org/10.18653/v1/2020.coling-main.506)]
- [130] Jiang F, Fan YX, Chu XM, Li PF, Zhu QM, Kong F. Hierarchical macro discourse parsing based on topic segmentation. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence (AAAI). AAAI, 2021. 13152–13160.
- [131] Chu XM, Zhu QM, Zhou GD. Discourse primary-secondary relationships in natural language processing. Chinese Journal of Computers, 2017, 40(4): 842–860 (in Chinese with English abstract). [doi: [10.11897/SP.J.1016.2017.00842](https://doi.org/10.11897/SP.J.1016.2017.00842)]
- [132] Jiang F, Chu XM, Xu S, Li PF, Zhu QM. A macro discourse primary and secondary relation recognition method based on topic similarity. Journal of Chinese Information Processing, 2018, 32(1): 43–50 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2018.01.006](https://doi.org/10.3969/j.issn.1003-0077.2018.01.006)]
- [133] Sun ZH, Zhou Y, Zhu QM, Jiang F, Li PF. Recognizing Chinese macro discourse nuclearity based on discourse topic. Journal of Chinese Information Processing, 2020, 34(12): 30–38 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2020.12.006](https://doi.org/10.3969/j.issn.1003-0077.2020.12.006)]
- [134] Zhou Y, Chu XM, Zhu QM, Jiang F, Li PF. Macro discourse relation classification based on macro semantics representation. Journal of Chinese Information Processing, 2019, 33(3): 1–7, 24 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2019.03.001](https://doi.org/10.3969/j.issn.1003-0077.2019.03.001)]
- [135] Jiang F, Li PF, Zhu QM. Joint modeling of recognizing macro Chinese discourse nuclearity and relation based on structure and topic gated semantic network. In: Proc. of the 8th CCF Int'l Conf. on Natural Language Processing and Chinese Computing (NLPCC). Dunhuang: Springer, 2019. 276–286. [doi: [10.1007/978-3-030-32236-6\\_24](https://doi.org/10.1007/978-3-030-32236-6_24)]
- [136] Sun ZH, Jiang F, Li PF, Zhu QM. Macro discourse relation recognition via discourse argument pair graph. In: Proc. of the 9th CCF Int'l Conf. on Natural Language Processing and Chinese Computing (NLPCC). Zhengzhou: Springer, 2020. 108–119. [doi: [10.1007/978-3-030-60457-8\\_9](https://doi.org/10.1007/978-3-030-60457-8_9)]
- [137] Jiang F, Li PF, Zhu QM. Recognizing Chinese discourse relations based on multi-perspective and hierarchical modeling. In: Proc. of the 2021 Int'l Joint Conf. on Neural Networks (IJCNN). Shenzhen: IEEE, 2021. 1–8. [doi: [10.1109/IJCNN52387.2021.9534273](https://doi.org/10.1109/IJCNN52387.2021.9534273)]
- [138] Tian WH, Gao YQ, Huang HW, Li ZW, Zhang ZY. Implicit discourse relation analysis based on multi-task Bi-LSTM. Journal of Chinese Information Processing, 2019, 33(5): 47–53 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2019.05.006](https://doi.org/10.3969/j.issn.1003-0077.2019.05.006)]
- [139] Tang YT, Li YB, Liu L, Yu ZH, Chen L. Feature learning by distant supervision for fine-grained implicit discourse relation identification. Acta Scientiarum Naturalium Universitatis Pekinensis, 2019, 55(1): 91–97 (in Chinese with English abstract). [doi: [10.13209/j.0479-8023.2018.060](https://doi.org/10.13209/j.0479-8023.2018.060)]
- [140] Guo FY, He RF, Dang JW. Implicit discourse relation recognition via a BiLSTM-CNN architecture with dynamic chunk-based max pooling. IEEE Access, 2019, 7: 169281–169292. [doi: [10.1109/ACCESS.2019.2954988](https://doi.org/10.1109/ACCESS.2019.2954988)]
- [141] Qin LH, Zhang ZS, Zhao H. Shallow discourse parsing using convolutional neural network. In: Proc. of the 2016 CoNLL-16 Shared Task. Berlin: ACL, 2016. 70–77. [doi: [10.18653/v1/K16-2010](https://doi.org/10.18653/v1/K16-2010)]
- [142] Liu Y, Zhang JJ, Zong CQ. Memory augmented attention model for Chinese implicit discourse relation recognition. In: Proc. of the 16th China National Conf. on Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Nanjing: Springer, 2017. 411–423. [doi: [10.1007/978-3-319-69005-6\\_34](https://doi.org/10.1007/978-3-319-69005-6_34)]
- [143] Rönnqvist S, Schenk N, Chiarcos C. A recurrent neural model with attention for the recognition of Chinese implicit discourse relations. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver: ACL, 2017. 256–262. [doi: [10.18653/v1/P17-2040](https://doi.org/10.18653/v1/P17-2040)]

- [144] Kuyten P, Bollegala D, Hollerit B, Prendinger H, Aizawa K. A discourse search engine based on rhetorical structure theory. In: Proc. of the 37th European Conf. on Information Retrieval. Vienna: Springer, 2015. 80–91. [doi: [10.1007/978-3-319-16354-3\\_10](https://doi.org/10.1007/978-3-319-16354-3_10)]
- [145] Sporleder C, Lapata M. Discourse chunking and its application to sentence compression. In: Proc. of the 2005 Conf. on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver: ACL, 2005. 257–264. [doi: [10.3115/1220575.1220608](https://doi.org/10.3115/1220575.1220608)]
- [146] Hewett F, Rane RP, Harlacher N, Stede M. The utility of discourse parsing features for predicting argumentation structure. In: Proc. of the 6th Workshop on Argument Mining. Florence: ACL, 2019. 98–103. [doi: [10.18653/v1/W19-4512](https://doi.org/10.18653/v1/W19-4512)]
- [147] Narasimhan K, Barzilay R. Machine comprehension with discourse relations. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing (ACL). Beijing: ACL, 2015. 1253–1262. [doi: [10.3115/v1/P15-1121](https://doi.org/10.3115/v1/P15-1121)]
- [148] Gao YF, Wu CS, Li JJ, Joty S, Hoi SCH, Xiong CM, King I, Lyu M. Discern: Discourse-aware entailment reasoning network for conversational machine reading. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2020. 2439–2449. [doi: [10.18653/v1/2020.emnlp-main.191](https://doi.org/10.18653/v1/2020.emnlp-main.191)]
- [149] Ji YF, Smith NA. Neural discourse structure for text categorization. In: Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (ACL). Vancouver: ACL, 2017. 996–1005. [doi: [10.18653/v1/P17-1092](https://doi.org/10.18653/v1/P17-1092)]
- [150] Marcu D. From local to global coherence: A bottom-up approach to text planning. In: Proc. of the 14th National Conf. on Artificial Intelligence and 9th Innovative Applications of Artificial Intelligence Conf. Providence: AAAI, 1997. 629–635.
- [151] Forsbom E. Rhetorical structure theory in natural language generation. Téléchargé le, 2005, 10(8): 2008.
- [152] Cojocaru DA, Trausan-Matu S. Text generation starting from an Ontology. In: Proc. of the 12th Romanian Human-computer Interaction Conf. Bucharest, 2015. 55–60.
- [153] Galitsky B, Ilvovsky D. Chatbot with a discourse structure-driven dialogue management. In: Proc. of the Software Demonstrations of the 15th Conf. of the European Chapter of the Association for Computational Linguistics (EACL). Valencia: ACL, 2017. 87–90. [doi: [10.18653/V1/E17-3022](https://doi.org/10.18653/V1/E17-3022)]
- [154] Liu MF, Qi QS, Hu HJ. A CNN Approach to football news generation based on discourse structure. Journal of Chinese Information Processing, 2019, 33(4): 101–108 (in Chinese with English abstract). [doi: [10.3969/j.issn.1003-0077.2019.04.012](https://doi.org/10.3969/j.issn.1003-0077.2019.04.012)]
- [155] Ko WJ, Li JY. Assessing discourse relations in language generation from GPT-2. In: Proc. of the 13th Int'l Conf. on Natural Language Generation. Dublin: ACL, 2020. 15–18.
- [156] Koto F, Lau J H, Baldwin T. Discourse probing of pretrained language models. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). ACL, 2021. 3849–3864. [doi: [10.18653/v1/2021.naacl-main.301](https://doi.org/10.18653/v1/2021.naacl-main.301)]
- [157] Akula AR, Zhu SC. Visual discourse parsing. arXiv:1903.02252, 2019.

#### 附中文参考文献:

- [2] 李艳翠. 汉语篇章结构表示体系及资源构建研究 [博士学位论文]. 苏州: 苏州大学, 2015.
- [13] 张迎, 王中卿, 王红玲. 基于篇章主次关系的单文档抽取式摘要方法研究. 中文信息学报, 2019, 33(8): 67–76. [doi: [10.3969/j.issn.1003-0077.2019.08.009](https://doi.org/10.3969/j.issn.1003-0077.2019.08.009)]
- [17] 付颖, 王红玲, 王中卿. 基于宏观篇章结构的科技论文摘要模型. 计算机应用, 2021, 41(10): 2864–2870. [doi: [10.11772/j.issn.1001-9081.2020121945](https://doi.org/10.11772/j.issn.1001-9081.2020121945)]
- [20] 亢晓勉, 宗成庆. 融合篇章结构位置编码的神经机器翻译. 智能科学与技术学报, 2020, 2(2): 144–152. [doi: [10.11959/j.issn.2096-6652.202016](https://doi.org/10.11959/j.issn.2096-6652.202016)]
- [32] 蒋峰. 汉语宏观篇章资源建设与结构分析方法研究 [硕士学位论文]. 苏州: 苏州大学, 2019. [doi: [10.27351/d.cnki.gszzhu.2019.000756](https://doi.org/10.27351/d.cnki.gszzhu.2019.000756)]
- [33] 张牧宇, 秦兵, 刘挺. 中文篇章级句间语义关系体系及标注. 中文信息学报, 2014, 28(2): 28–36. [doi: [10.3969/j.issn.1003-0077.2014.02.004](https://doi.org/10.3969/j.issn.1003-0077.2014.02.004)]
- [40] 严为绒, 徐扬, 朱珊珊, 洪宇, 姚建民, 朱巧明. 篇章关系分析研究综述. 中文信息学报, 2016, 30(4): 1–11.
- [46] 孔芳, 王红玲, 周国栋. 汉语篇章理解研究综述. 软件学报, 2019, 30(7): 2052–2072. <http://www.jos.org.cn/1000-9825/5834.htm> [doi: [10.13328/j.cnki.jos.005834](https://doi.org/10.13328/j.cnki.jos.005834)]
- [50] 吴为章, 田小琳. 汉语句群. 北京: 商务印书馆, 2000.
- [51] 邢福义. 汉语复句研究. 北京: 商务印书馆, 2001.

- [52] 郝晓燕, 刘伟, 李茹, 刘开瑛. 汉语框架语义知识库及软件描述体系. 中文信息学报, 2007, 21(5): 96–100, 138. [doi: 10.3969/j.issn.1003-0077.2007.05.018]
- [53] 褚晓敏. 宏观篇章结构表示体系、资源建设和计算模型研究 [博士学位论文]. 苏州: 苏州大学, 2018. [doi: 10.27351/d.cnki.gszzhu.2018.000122]
- [54] 褚晓敏, 奚雪峰, 蒋峰, 徐昇, 朱巧明, 周国栋. 宏观篇章结构表示体系和语料建设. 软件学报, 2020, 31(2): 321–343. <http://www.jos.org.cn/1000-9825/5868.htm> [doi: 10.13328/j.cnki.jos.005868]
- [114] 孙成, 孔芳. 基于转移的中文篇章结构解析研究. 中文信息学报, 2018, 32(12): 48–56. [doi: 10.3969/j.issn.1003-0077.2018.12.006]
- [116] 王体爽. 中文微观篇章分析方法研究 [硕士学位论文]. 苏州: 苏州大学, 2020. [doi: 10.27351/d.cnki.gszzhu.2020.001991]
- [120] 徐昇. 中文微观篇章主次和关系识别方法研究 [硕士学位论文]. 苏州: 苏州大学, 2019. [doi: 10.27351/d.cnki.gszzhu.2019.000679]
- [121] 王体爽, 李培峰, 朱巧明. 基于门控记忆网络的汉语篇章主次关系识别方法. 中文信息学报, 2019, 33(5): 39–46. [doi: 10.3969/j.issn.1003-0077.2019.05.005]
- [122] 徐昇, 王体爽, 李培峰, 朱巧明. 运用多层注意力神经网络识别中文隐式篇章关系. 中文信息学报, 2019, 33(8): 12–19, 35. [doi: 10.3969/j.issn.1003-0077.2019.08.002]
- [127] 周懿. 汉语宏观篇章树自动构建方法研究 [硕士学位论文]. 苏州: 苏州大学, 2020. [doi: 10.27351/d.cnki.gszzhu.2020.001566]
- [131] 褚晓敏, 朱巧明, 周国栋. 自然语言处理中的篇章主次关系研究. 计算机学报, 2017, 40(4): 842–860. [doi: 10.11897/SP.J.1016.2017.00842]
- [132] 蒋峰, 褚晓敏, 徐昇, 李培峰, 朱巧明. 基于主题相似度的宏观篇章主次关系识别方法. 中文信息学报, 2018, 32(1): 43–50. [doi: 10.3969/j.issn.1003-0077.2018.01.006]
- [133] 孙振华, 周懿, 朱巧明, 蒋峰, 李培峰. 基于篇章主题的中文宏观篇章主次关系识别方法. 中文信息学报, 2020, 34(12): 30–38. [doi: 10.3969/j.issn.1003-0077.2020.12.006]
- [134] 周懿, 褚晓敏, 朱巧明, 蒋峰, 李培峰. 基于宏观语义表示的宏观篇章关系识别方法. 中文信息学报, 2019, 33(3): 1–7, 24. [doi: 10.3969/j.issn.1003-0077.2019.03.001]
- [138] 田文洪, 高印权, 黄厚文, 黎在万, 张朝阳. 基于多任务双向长短时记忆网络的隐式句间关系分析. 中文信息学报, 2019, 33(5): 47–53. [doi: 10.3969/j.issn.1003-0077.2019.05.006]
- [139] 唐裕婷, 李艳斌, 刘露, 于中华, 陈黎. 面向细粒度隐式篇章关系识别的远距离监督特征学习算法. 北京大学学报(自然科学版), 2019, 55(1): 91–97. [doi: 10.13209/j.0479-8023.2018.060]
- [154] 刘茂福, 齐乔松, 胡慧君. 基于卷积神经网络与篇章结构的足球新闻自动生成方法. 中文信息学报, 2019, 33(4): 101–108. [doi: 10.3969/j.issn.1003-0077.2019.04.012]



蒋峰(1994—), 男, 博士生, CCF 学生会员, 主要研究领域为自然语言处理, 篇章分析.



李培峰(1971—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 机器学习.



范亚鑫(1997—), 男, 博士生, CCF 学生会员, 主要研究领域为自然语言处理, 篇章分析.



朱巧明(1963—), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为中文信息处理, Web 信息处理.



褚晓敏(1981—), 女, 博士, CCF 专业会员, 主要研究领域为自然语言处理, 篇章分析.