# 自然语言处理中的篇章主次关系研究

褚晓敏 朱巧明 周国栋

(苏州大学计算机科学与技术学院 江苏 苏州 215006)

(苏州大学自然语言处理实验室 江苏 苏州 215006)

**摘 要** 篇章结构分析特别是篇章主次关系研究是自然语言处理领域的一个重要研究方向.篇章主次关系的分析,有助于理解篇章的结构和语义,并为自然语言处理的应用(例如自动文摘、主题抽取和问答系统等)提供有力的支持.然而,目前篇章主次关系分析却是篇章结构分析的一个瓶颈.已有研究一般将篇章主次关系分析看作篇章修辞结构分析中的一个辅助环节,忽略了其在篇章结构分析中的重要性.因此,文中将篇章主次关系提升到篇章结构分析的核心地位,将它从篇章修辞结构分析中分离出来作为一个独立的任务进行研究.首先,探讨了什么是篇章主次关系、如何判别篇章主次关系以及为什么要研究篇章主次关系;其次,分别从两个角度(微观、宏观)和三个方面(理论体系、语料资源和计算模型)详细阐述了篇章主次关系的研究现状;再次,分析了篇章主次关系研究存在的问题,并提出我们的基本研究思路;最后,归纳出篇章主次关系未来的一些研究方向.

**关键词** 自然语言处理;篇章结构分析;篇章主次关系;宏观主次关系;微观主次关系;社会媒体

**中图法分类号** TP18 **DOI号** 10.11897/SP.J.1016.2017.00842

# Discourse Primary-Secondary Relationships in Natural Language Processing

CHU Xiao-Min ZHU Qiao-Ming ZHOU Guo-Dong

(*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215006)

(*Natural Language Processing Lab, Soochow University, Suzhou, Jiangsu* 215006)

**Abstract** Discourse structure analysis, especially recognizing the primary-secondary relationship in discourse structures is an important research topic in natural language processing. Recognition of discourse primary-secondary relationship not only helps to understand the discourse structure and semantics, but also provides strong support for deep applications of natural language processing, such as summarization, topic extraction, question answering, etc. However, discourse primary-secondary relationship recognition is bottleneck of discourse structure analysis in current discourse researches. Most existing research views the recognition of primary-secondary relationship as a dispensable component attached to the analysis of the rhetorical structure, totally ignoring the important role of primary-secondary relationship played in document level discourse structure analysis. Nevertheless, this paper regards the recognition of primary-secondary relationship as an independent task from the discourse rhetorical structure analysis, illustrating its critical role in discourse structure analysis. First, the paper discusses the definition of primary-secondary relationship, how to determine the primary-secondary relationship and its importance in discourse structure analysis. Second, the paper summarizes the research status of recognizing the primary-

收稿日期:2016-05-11;在线出版日期:2016-09-23.本课题得到国家自然科学基金(61272260)、教育部中国移动科研基金(MCM20150602)、江苏省科技计划(BK20151222)资助.**褚晓敏**,女,1981年生,博士研究生,主要研究方向为自然语言处理、篇章分析.E-mail:64262009@qq.com.**朱巧明**(通信作者),男,1963年生,博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为自然语言处理和Web信息处理.E-mail:qmzhu@suda.edu.cn.**周国栋**,男,1966年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为自然语言处理、信息抽取、统计机器翻译和机器学习.

secondary relationship in discourse structure from both macro-level and micro-level, and from three aspects, i. e., theory system, corpus resource, computing model. Moreover, this paper presents our several proposals targeting at the key issues in research on the primary-secondary relationship. Last but not least, we present several directions of future work related to the primary-secondary relationship.

**Keywords**　natural language processing; discourse structure analysis; discourse primary-secondary relationships; macro primary-secondary relationships; micro primary-secondary relationships; social media

## 1　引　言

近年来，自然语言处理领域的研究重点逐步从浅层次的词汇、句法分析延伸到了深层次的语义理解. 具体而言，也就是逐渐从传统的词汇、句法和浅层语义角色标注等，深入到语义连贯性和结构衔接性的研究上来. 因此，自然语言处理研究的文本颗粒度，也从单个词、短语和句子，延伸至句群、段落和篇章. 篇章分析在自然语言处理和计算语言学研究领域得到了前所未有的发展，成为最活跃的研究方向之一.

篇章指篇幅与章节，是自然语言文本理解的研究对象，是通过语义关联和结构化组织形成的自然语言文本. de Beaugrande 和 Dressler[1] 在 1981 年的论著中论述了篇章的 7 个基本特征，分别是衔接性(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(intertextualinty). 这 7 个特征涉及篇章语言学所关注的 3 个方面：篇章、参与者和语境. 其中，衔接性和连贯性是对篇章自身的语言形式和语义而言的；意图性和可接受性是分别针对篇章的产生者和接受者，即篇章的参与者而言的；信息性和情景性涉及篇章的信息程度和与情景之间的关系，跨篇章性则描述了篇章与其他篇章之间的关系，即这三个特征是与篇章的语境紧密相关的.

篇章的衔接性和连贯性是篇章的两个最基本特征，也是篇章研究和分析的核心问题. 衔接是篇章表层的形式上衔接，主要表现为词汇、短语或子句之间的联接，即篇章表层结构上的各语言成分之间的语法或语义的关系. 而连贯是深层的语义联接，主要表现为句子、复句或句群之间的关联，即篇章深层的意义上连贯. 从本质上来看，衔接性和连贯性分别从形式和语义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. 有了衔接性和连贯性，整个篇章才更完整.

篇章结构分析的研究成果在自然语言处理领域应用广泛，例如在统计机器翻译(Statistical Machine Translation)[2-4]、自动文摘(Text Summarization)[5-7]、自动问答系统(Question Answering System)[8]、信息抽取(Information Extraction)[9-10] 和情感分析(Sentiment Analysis)[11-13] 等领域都取得了一定的研究成果. 这些应用研究的成果表明，对文本进行更深层次的挖掘，包括对结构和语义信息的深入分析，有助于在当前主要基于统计的自然语言处理研究方法上取得新的进展.

篇章主次关系的研究对象是句子、句群和段落之间的语义关联和它们的重要性关系，表现的是连贯性这个篇章基本特征. 作为篇章结构分析的一个重要环节，篇章主次关系研究的目的是分析篇章的主要内容和次要内容，进而理解篇章主题思想、展开思路和主要内容.

篇章局部的重要性的判断离不开篇章全局，只有在整体上把握篇章主旨才能更好地分析篇章主次关系，而研究篇章主次关系，也能更好的认识和理解篇章的中心主题和展开思路，更有效的挖掘篇章的宏观主题和篇章各部分之间的语义关联，并为自然语言处理的相关应用，例如主题抽取、自动文摘和问答系统等提供强有力的支持.

本文首先介绍什么是篇章主次关系，分析为什么研究篇章主次关系；其次，从微观和宏观两个角度介绍目前篇章主次关系在理论体系、语料资源和计算模型等三个方面国内外学者的主要研究成果和代表性工作；然后，分析篇章主次关系研究中存在的问题，并提出对应的研究策略；最后，分析并讨论篇章主次关系未来的一些研究方向.

## 2 篇章主次关系

### 2.1 什么是篇章主次关系

篇章主次关系简单地说,就是一个篇章内部的主要内容和次要内容之间的关系,或者篇章与篇章之间主要方面与次要方面的关系.其中,主要内容是指篇章中居于支配地位、起决定作用的部分,而次要内容是指篇章中居于辅助地位、不起决定作用的部分.

一个篇章关系一般包含两个篇章单位,这两个篇章单位同属一个关系层,如果其中一个篇章单位能够概括它所在关系层主旨和内容,能代表其所在关系层与外界发生关系,则这种关系为单核关系;如果两个篇章单位同等重要,则这种关系为多核关系.例如,在陈述-举例关系连接的两个篇章单位中,一个是陈述项,一个是举例项,举例项是为陈述项服务的,因此陈述项是该篇章关系的核心,陈述-举例关系是单核关系;在并列关系中,篇章单位可以有两个或多个,并列关系的核心可能会由一个或多个篇章单位来充当,即并列关系可能是单核关系,也可能是多核关系.

在微观角度,篇章主次关系表现为,句子与句子之间、句群与句群之间的主要和次要的关系;在宏观角度,篇章主次关系表现为,段落与段落、章节与章节之间的主要和次要的关系.下面以具体的例子来进一步说明篇章主次关系的含义.

我们以宾州汉语树库(Penn Chinese TreeBank,CTB)[14]中的3个具体的例子来说明微观主次关系的含义(出自 chtb_0001,《上海浦东开发与法制建设同步》,例子中 a、b、c、d、e、f、g 为基本篇章单位的编号):

**例 1**. a 尽管浦东新区制定的法规性文件有些比较"粗",b 有些还只是暂行规定,c 有待在实践中逐步完善,|d 但这种法制紧跟经济和社会活动的做法,受到了国内外投资者的好评,e 他们认为,到浦东新区投资办事有章法,f 讲规矩,g 利益能得到保障.

这个例子中,以"|"划分前后两个篇章单位,其中后者表达的语义信息"受到了国内外投资者的好评"比前者"法规性文件有些比较'粗'"更重要,因此本例是一个单核关系,核心在后,属于"让步关系".

**例 2**. d 但这种法制紧跟经济和社会活动的做法,受到了国内外投资者的好评,|e 他们认为,到浦东新区投资办事有章法,f 讲规矩,g 利益能得到保障.

这个例子中,以"|"划分前后两个篇章单位,后一个篇章单位进一步阐述了前一个篇章单位中提到的"国内外投资者的好评",因此本例是一个单核关系,核心在前,属于"解说关系".

**例 3**. e 他们认为,到浦东新区投资办事有章法,|f 讲规矩,|g 利益能得到保障.

这个例子中,以"|"划分3个篇章单位,这3个篇章单位之间语义表达上并无主次之分,因此本例是一个多核关系,属于"并列关系".

这3个例子在篇章关系分析的同时也展现了篇章主次关系自顶向下,逐层分析的过程,篇章结构分析完成后,该篇章可以用一个完整的依存树结构来表示,如图1的连接依存树所示.在该依存树结构中,叶子节点为基本篇章单位,非叶子节点为篇章关系.带箭头的分支指向篇章关系中的主要部分,不带箭头的分支指向次要部分.

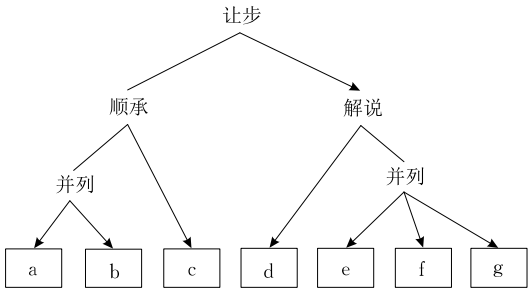



图 1 例 1~3 的连接依存树表示

在图1的依存树结构中,依箭头指向可以发现,叶子节点 d 是该段最主要的篇章单元,在抽取文本摘要时可以将此句作为信息抽取重点.结合该篇章的话题,抽取"上海浦东新区这种法制紧跟经济和社会活动的做法,受到了国内外投资者的好评"作为该篇章的文本摘要,是非常贴切和恰当的.由此可以看出,篇章主次关系为文本摘要应用研究提供了有效、便捷的手段.

我们再以 CTB 中的另一个例子来说明篇章宏观主次关系的含义.

**例 4**. 在 chtb_0019《宁波保税区建设成就显著》这个篇章中,有如下段落.

a 新华社宁波二月七日电(记者胡宏伟、柴骥程)总面积二点三平方公里的宁波保税区,经过三年建设,已取得丰硕成果.

b 宁波保税区是中国十三个保税区之一,于一九九二年经国务院批准设立.目前,保税区的各项功能已初具规模,开发水平在中国各保税区中名列前茅.

c 据统计，至去年年底，宁波保税区累计完成进出口贸易额八点一二亿美元，仅去年一年通过保税区海关的进出口贸易额就达三点六五亿美元. 目前，区内已有十个保税仓库，仓储面积达八万多平方米；仅去年一年，区内储有货物就达二十六点二七亿元人民币.

d 随着从今年四月开始中国对保税区外有关特殊政策的调整，保税区免证、免税，保税政策的稳定性优势显得更为明显，国内外一大批实业加工项目相继在区内落户. 到去年十二月底，区内已累计设立企业一千六百一十四家，总投资达十二亿美元，其中外商投资企业二百六十家，实际利用外资一点一三亿美元. 另外，众多国内企业也通过保税区与国际市场接轨.

e 为了在运行机制上与保税区相配套，宁波保税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，一次性办理. 同时，保税区大力抓好区内信息高速公路的网络体系建设，为实现现代化管理创造良好的配套条件.

图 2 显示了例 4 的宏观篇章结构. 在该篇章结构树中，叶子节点表示篇章中的段落，内部节点表示段落之间的关系，箭头指向表示篇章主次关系. 在这个篇章中，段落 b 介绍了“宁波保税区”的批准设立情况和目前的发展情况，是段落 a 中“宁波保税区已取得丰硕成果”这个事件的背景，因此，段落 a 与段落 b 之间构成了“事件-背景关系”，描述事件的段落 a 是主要内容，描述背景的段落 b 是次要内容. 段落 c、d 和 e 分别从三个方面阐述了段落 a 中提及的“已取得丰硕成果”，这三个段落之间是“并列关系”，而它们构成的整体解说了由段落 a、b 构成的整体，形成“解说关系”，因此由段落 a、b 构成的整体更重要.

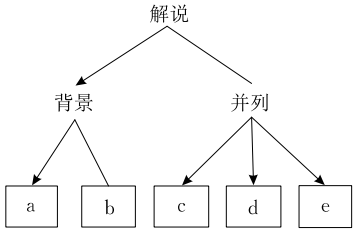



图 2　例 4 的宏观篇章树结构表示

## 2.2　如何判别篇章主次关系

在篇章中判断篇章单位的重要性离不开篇章全局，单纯看一个篇章关系中的两个篇章单位往往难于直接判断哪个重要. 比如因果关系，可能是“结果”重要，也可能是“原因”重要，究竟哪个更重要需要根据篇章的全局结构和语义做出判断. 另外，篇章单位的主次区分还与语序有关，同一篇章关系中，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，可能导致篇章主次关系的变化. 下面以一组“因果关系”的实际例子对主次关系的判别加以说明.

**例 5**. 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.

这个例子中，以“|”划分前后两个篇章单位，表达了因为“采购服务中心一开始就比较规范”，所以“成交药品一亿多元，没有发现一例回扣”的含义，强调了事件的“原因”，因此在该例中主要篇章单位是前者，原因项.

**例 6**. 随着中国对大运河的整治，运河航道状况得到极大改善，|许多企业纷纷看好这条“黄金水道”，积极在此投资建厂，沿河企业星罗棋布.

这个例子中，以“|”划分前后两个篇章单位，表达了因为“运河航道状况得到极大改善”，所以“许多企业积极在此投资建厂”的含义，强调了事件的“结果”，因此在该例中主要篇章单位是后者，结果项.

一般来说，在篇章中，每一个关系层，总有一个子句、句子或段落，能够概括该关系层的内容，与上下文联系更加紧密，并且更紧扣篇章的主题，这样的子句、句子或段落就是这个关系层的主要部分. 根据篇章主次关系的定义和上述特点，归纳篇章主次关系的 3 个基本判断标准：(1) 是否能概括它所在关系层的内容和意图，是否能代表所在篇章关系层与外界发生关系；(2) 与上下文的联系程度是否紧密；(3) 是否符合篇章主旨，与篇章主题的关联程度是否紧密. 符合这 3 个标准的篇章单位即为所在关系层的主要篇章单位.

在例 2 中，篇章单位 e、f 和 g，都是对 d“受到了国内外投资者的好评”内容的解释说明，即 d 更能代表这个篇章关系层与上下文发生关系；同时，在 a、b、c 组成的整个篇章单位和 d、e、f、g 组成的整个篇章单位之间，构成“让步关系”，核心在后，因此篇章单位 d 在整个篇章中表达最为重要的含义，更符合篇章的主旨，更符合篇章想要表达的对“浦东新区制定的法规性文件”这个事件的正面评价的意图. 因此，根据篇章主次关系的判断标准，篇章单位 d 是整个篇章中最主要的部分，表达最重要的篇章语义. 在图 1 的连接依存树结构中，也可以依据箭头的指向，

寻找到最重要的篇章单位 d.

在例 4 中,篇章单位 c、d、e 分别从宁波保税区的贸易额与仓储建设、政策调整和信息化建设这 3 个方面解说了宁波保税区“已取得丰硕成果”,它们之间是并列关系,没有主次之分,因此 c、d 和 e 同等重要. 全篇中,段落 a 最能表达“宁波保税区建设成就显著”的篇章主题(也是该篇章的标题),因此段落 a 是这个篇章中最重要的篇章单位.

### 2.3 为什么要研究篇章主次关系

根据 2.1 章节对什么是篇章主次关系的阐述以及例 1～4 的篇章结构分析过程,可以了解到,篇章主次关系对篇章结构的构建、篇章语义的理解和篇章主题的识别都有很大的帮助. 也就是说,研究篇章主次关系,有利于分析篇章的中心主题,有利于分析和构建篇章的整体结构,有利于理解篇章的语义信息.

例如 2.1 章节中的例 4,根据篇章结构和篇章主次关系可知,段落 a 是全文最重要的段落,并且结合文章的标题,可进一步验证段落 a 符合篇章的主旨,与篇章主题的相关程度更高. 因此,可以直接抽取段落 a 作为全文的摘要.

尽管篇章主次关系的识别和分析对篇章分析有着重要的作用,已有的研究却一般只是将篇章主次关系分析看作篇章修辞结构分析中的一个辅助环节,并没有给予足够的重视.

因此,考虑到篇章主次关系的重要性,以及在针对篇章主次关系研究意义的分析的基础上,本文将把篇章主次关系作为独立的研究对象进行研究,通过深入探索篇章微观和宏观主次关系以及它们之间的交互作用,建立理论体系、建设语料资源、构建计算模型,将有效的提高篇章主次关系的识别性能,进而更准确地分析篇章结构,更深入地理解篇章语义,相应的研究成果可应用到自然语言处理的相关应用中.

## 3 研究现状分析

近年来,随着篇章级语料库(尤其是 PDTB 和 RST-DT)的不断建设,以及一些国际学术评测(例如 CoNLL2015、CoNLL2016 等)的大力推动,篇章结构分析发展迅速,成为自然语言处理领域最活跃的研究方向之一.

篇章结构的研究分析可分为微观和宏观两个角度. 微观篇章结构是指篇章中的一个句子内部的结构或两个连续的句子之间的结构,宏观篇章结构是指更高层次的结构,表现为句群、段落和章节之间的结构.

微观角度的篇章结构理论主要包括浅层衔接理论[15]、Hobbs 模型[16-17]、修辞结构理论[18-22]、宾州篇章树库理论[14,23-25]、意图结构理论[26-27]、信息结构理论[28]、主位述位理论[29]、D-LTAG 理论[30]、句群理论[31]、复句理论[32-33]和基于连接依存树的篇章结构理论[34-35]等. 宏观角度的篇章结构理论则相对较少,主要包括篇章模式[36]、超主位理论[37]和篇章宏观结构理论[38-42]等.

篇章主次关系作为篇章结构的一个重要组成部分,一直融合在篇章结构的研究中. 其理论体系的研究,最早可追溯到 1975 年,Grimes[43]在他的研究中首次提出了衔接关系的论元有主次之分. 之后在微观角度和宏观角度各有一些理论体系逐渐构建起来. 例如在微观角度方面,Mann 和 Thompson[18]在他们创立的修辞结构理论(Rhetorical Structure Theory,RST)中指出,修辞关系连接的两个或多个篇章单位之间存在主次之别,形成“核心-卫星”结构,其中,能够表达篇章主要信息的篇章单位称作“核心(Nucleus)”,而表达次要信息,对“核心”起到支撑作用的篇章单位称作“卫星(Satellite)”;宏观角度,van Dijk[38]提出了篇章的宏观结构理论,他认为宏观结构是篇章整体上的高层次的结构,表达篇章的主题思想. 语料资源和计算模型的研究,在微观角度得到了一定的发展,例如修辞结构篇章树库是目前篇章主次分析的重要资源,相关的计算模型研究也较多;反观宏观角度,相关研究则相当匮乏,既没有语料资源构建的实践,也没有计算模型的相关研究;更没有研究者将微观和宏观统一为一个整体,并基于统一的整体构建相应的理论表示体系、语料资源和计算模型.

综上所述,已有的篇章主次关系的研究可以分为微观和宏观两个角度. 现有研究涉及微观主次关系的理论、资源和模型相对丰富,而宏观主次关系还停留在理论层面,缺乏相应的资源和模型.

### 3.1 篇章微观主次关系

#### 3.1.1 理论体系

涉及篇章微观主次关系的篇章结构理论研究主要有:浅层衔接理论、修辞结构理论、意图结构理论和基于连接依存树的篇章结构理论. 其他的一些篇章结构理论,如 Hobbs 模型、宾州篇章树库理论、句

群理论和复句理论等未涉及篇章主次关系.本章节将简要介绍涉及篇章微观主次关系的理论体系研究现状.

(1)浅层衔接理论

Halliday[15]的浅层衔接理论是最早研究篇章衔接关系的理论体系.浅层衔接理论指出,“当篇章中的某个成分的解释依赖于篇章中另一个成分的解释时,这两个成分之间就产生了衔接关系”;衔接方式通常分为语法衔接和词汇衔接两大类,其中语法衔接手段包括指称、省略、替代和(逻辑)连接,连接又划分为增补型(Additive)、转折型(Contrastive)、原因型(Causal)和时间型(Temperal)四类,词汇衔接手段包括词汇的重复和搭配.

Grimes[43]在深化 Halliday 的浅层衔接理论时考虑非词汇化的命题关系,给出了更详细的衔接关系类别,此外,Grimes 首次提出了衔接关系的论元有主次之分,并明确的指出,并列(Paratactic)关系的论元同等重要,而主从(Hypotactic)关系的论元有主次之分.

(2)修辞结构理论

Mann 和 Thompson[18-22]在 20 世纪 80 年代创立了修辞结构理论(Rhetorical Structure Theory,RST).修辞结构理论是他们在系统功能理论的框架下,提出的关于篇章生成和分析的理论.三十多年来,在这一理论框架下,研究人员建设了语料资源、构建了计算模型和修辞结构理论对篇章结构分析的发展有重要的意义.修辞结构理论认为篇章各小句不是杂乱无章的堆放在一起的,而是由小句之间的修辞关系关联在一起,较小的功能块通过修辞关系组成更大的功能块.修辞结构理论从系统功能的角度出发,通过对句子与句子之间的修辞关系的研究,聚焦篇章的整体性和连贯性.修辞结构理论通过篇章修辞关系的描写,反应篇章生产者的交际意图和中心思想.

修辞结构理论共定义了 4 大类 25 小类修辞关系,例如环境(Circumstance)关系、目的(Purpose)关系、解答(Solutionhood)关系、对照(Antithesis)关系、阐述(Elaboration)关系、让步(Concession)关系、背景(Background)关系和评价(Evaluation)关系等.每个修辞关系可以连接两个或多个篇章单位,详细规定了在分析修辞结构时需要做出的各种具体判断.RST 运用一套专门的术语来描述篇章的关系结构:(1)文本范围(Span),指篇章结构中具有修辞结构和功能完整的文本跨度,两个独立的文本范围根据其在修辞关系中的地位,分为“核心文本范围”和“卫星文本范围”;(2)篇章单位(Discourse Unit),指构成文本范围的表层篇章单位,一般是子句;(3)关系定义(Relation Definition),两个文本范围之间修辞的定义和判断依据.包括两个方面:一是限制条件(Constraints),包括核心文本范围限制条件、卫星文本范围限制条件及这两种文本范围联合限制条件;二是效果(Effect),指作者运用该修辞关系所希望达到的效果及效果轨迹.

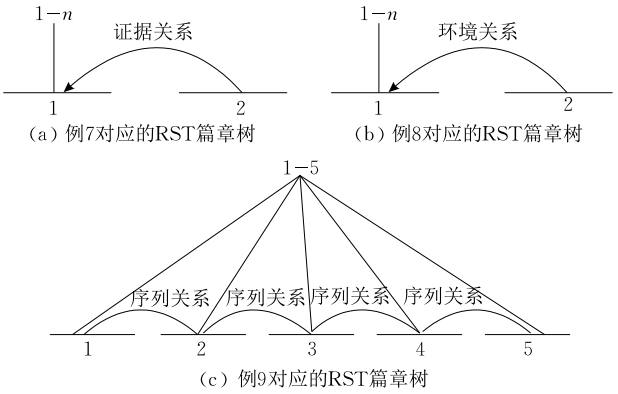
图 3 以证据(Evidence)关系为例展示修辞关系的定义.证据关系连接的两个文本范围中,核心文本范围陈述观点,卫星文本范围为之提供证据支持.使用证据关系所期望达到的效果是提高读者对核心文本范围的相信程度.

关系名称:证据关系
核心文本范围限制条件:读者可能认为核心文本范围的可相信程度不足
卫星文本范围限制条件:读者认为卫星文本范围的可相信程度较强
核心-卫星文本范围限制条件:通过对卫星文本范围的理解增强读者对核心文本范围相信程度
效果:增加读者对核心文本范围的相信程度
效果轨迹:核心文本范围

图 3　RST 中证据关系的定义

由这个定义,可以看出修辞关系连接的两个或多个文本范围并不是对称的,而是存在着主要和次要的区别,形成一种“核心-卫星”的结构,其中,“核心(Nucleus)”是主要文本范围,而“卫星(Satellite)”是次要文本范围,“卫星”文本范围为“核心”文本范围表达的信息服务,辅助读者了解、理解和相信“核心”文本范围所表达的内容.文本范围存在主次之分的修辞关系称为“单核(Mononuclear)”关系,例如背景关系(Background)、证据关系(Evidence)和目的关系(Purpose)等;文本范围之间没有主次之分的修辞关系称为“多核(Multi-nuclear)”关系,例如序列关系(Sequence)和对比关系(Contrast)等.

当两个以上的文本范围形成修辞关系时,就构成了“树”结构,即修辞结构树.在 RST 结构树中,叶子节点代表篇章单位(Discourse Unit),在连接篇章单位的弧线上标明文本间的修辞关系,水平线表示文本范围(Span),垂直线对应的文本范围为该修辞关系的“核心”.为清晰起见,我们举例加以说明,图 4 展示了例 7～9 对应的 RST 树分析的过程.




(a) 例7对应的RST篇章树 (b) 例8对应的RST篇章树

(c) 例9对应的RST篇章树

图 4 例 7～例 9 对应的 RST 篇章树

**例 7**. (1) They are having a party again next door; (2) I couldn't find a parking space.

在这个例子中,核心小句(1)陈述观点,卫星小句(2)为之提供证据,通过"I couldn't find a parking space"提高"They are having a party again next door"的可相信程度. 该例为"证据关系",其 RST 结构如图 4(a)所示.

**例 8**. (1) Probably the most extreme case of Visitors Fever I have ever witnessed was a few minutes ago; (2) when I visited relatives in the Midwest.

在这个例子中,卫星小句(2)为核心小句(1)设定了时间框架. 该例为"环境关系",其 RST 结构如图 4(b)所示.

**例 9**. (1) Peel oranges; (2) and slice crosswise; (3) Arrange in a bowl; (4) and sprinkle with rum and coconut; (5) Chill until ready to serve.

在这个例子中,这 5 个小句之由序列关系连接,表示这 5 个动作是顺序进行的,5 个小句都是核心小句. 该例为"序列关系",其 RST 结构如图 4(c)所示.

(3) 意图结构理论

Grosz 和 Sidner[26-27] 认为,篇章是具有意图的,因此篇章分析不能只考虑篇章的结构,还应理解篇章所要表达的意图,在他们提出的篇章结构中,包括 3 个方面,分别是语言结构(Linguistic structure)、意图结构(Intentional structure)和焦点状态(Attentional state).

根据 Grosz 和 Sidner 对篇章结构的定义,篇章意图(Discourse Purpose, DP)由篇章段意图(Discourse Segment Purpose, DSP)分解和表达,显示出篇章意图的层次性特点. 同一个意图层,如果 DSP1 有助于表达 DSP2,则 DSP2 占主导地位,并称之为支配关系(Dominance),支配关系与修辞结构理论中的"核心-卫星"结构相似,因此可以看作是主次关系在篇章意图层上的定义.

(4) 基于连接依存树的篇章结构理论

Li 等人[34-35] 在针对连接依存树[44] 和篇章结构分析研究[45-50] 的基础上,吸取了修辞结构理论的树形结构和篇章主次关系,参考了宾州篇章树库对连接词的处理方式,同时结合汉语复句和句群理论,提出了基于连接依存树的篇章结构理论(Chinese Discourse Structure Theory based on Connective-driven Dependency Tree),该理论对完整的篇章结构(包括篇章单位、连接词、篇章结构、篇章关系和篇章主次)进行了系统的定义和描述.

在基于连接依存树的篇章结构中,叶子节点表示基本篇章单位(Elementary Discourse Units, EDUs),内部节点为连接词(Connective),由连接词连接的基本篇章单位组合称为篇章单位(Discourse Units, DUs). 各子句之间通过连接词形成更高一级的篇章单位,层次组合直至形成一棵完整的篇章结构树. 连接词既可以表示篇章单位层次,也可以表示篇章单位之间的逻辑语义关系,一个连接词可以连接两个或多个篇章单位,篇章单位根据在篇章中的重要程度可分为主要篇章单位和次要篇章单位,主要篇章单位代表所在关系层整体与外界发生关系,而次要篇章单位辅助主要篇章单位意义的表达. 例 10 是一个汉语篇章,图 5 是根据基于连接依存树的篇章结构理论构建的篇章结构树.

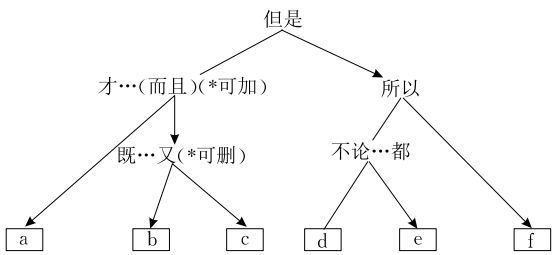



图 5 例 10 的篇章结构连接依存树表示

**例 10**. a 张三才 30 出头,b(而且)既没有什么学历,c 又没用多少新的工作经验,d 但是不论干什么,e 他都非常认真,f 所以,处长总是把一些重要的任务交给他.

图 5 中的字母标记的叶子节点表示基本篇章单位,内部节点为连接词(在显式篇章关系中为连接词,在隐式篇章关系中为添加的连接词),内部节点

所连接的子句称为篇章单位.各级篇章单位通过连接词组合后形成高一级篇章结构,进而形成完整的篇章结构树.连接词的层级地位可以反映篇章结构,连接词本身对应篇章关系.篇章单位通过连线的指向性可区分篇章单位的主次地位,带箭头的篇章单位为主要篇章单位,而不带箭头的则为次要篇章单位.

判断主要和次要篇章单位的时候需要根据考虑篇章单位能否概括它所在关系层的内容和意图,与上下文联系是否紧密以及与篇章主题的相关程度如何,在该例中通过篇章单位 d、e 对张三"非常认真"这个特点的描述,阐述了篇章表达意图"处长总是把一些重要的任务交给他"的原因,而篇章单位 a、b、c 形成的整体与篇章单位 d、e、f 形成的整体通过连接词"但是"连接,构成转折关系,显然,相对于"张三"的不足相比,作者更强调"他"的"认真",d、e、f 形成的整体是主要篇章单位,a、b、c 形成的整体是次要篇章单位.

3.1.2　语料资源

目前涉及到篇章主次关系语料资源并不多,其中英语的篇章语料库主要包括修辞结构篇章树库(RST-DT)和篇章图库(Discourse GraphBank)等.而汉语的篇章语料库建设相对较晚较少,主要包括借鉴 RST 标记的汉语篇章语料库(CJPL)和基于连接依存树的汉语篇章树库(CDTB).

(1)修辞结构篇章树库

修辞结构篇章树库 RST-DT(Rhetorical Structure Theory Discourse Treebank)[51]是以修辞结构理论(RST)为基础的篇章语料库,由美国南加利福尼亚大学和华盛顿国防部联合标注,2002 年由 LDC 发布.RST-DT 标注了 385 篇华尔街日报的文章,文章长度从 31 个词到 2124 个词不等,平均每篇文章 458.14 个词,语料总词数达到 176 000 个,标注的 EDU 总数为 21789 个.

RST-DT 的标注分为两大步骤,分别是 EDUs 的划分和篇章结构的构建.RST-DT 定义子句为篇章的 EDUs,并利用词汇和句法线索来辅助 EDUs 的划分.邻接的 EDUs 通过篇章关系形成篇章结构,修辞结构理论给出了 4 大类 25 小类修辞关系,并明确指出这是一个开放关系集,这意味着读者可以扩展修辞关系集,定义其他修辞关系类型.RST-DT 标注了 53 种单核关系和 25 种多核关系,分成 16 个组别,每组具有相同的修辞功能[18-19].

(2)篇章图库

Wolf 和 Gibson[52]认为基于树结构的篇章结构表示存在一定的局限,不能有效表示句子之间存在多种篇章关系的情况.因此他们提出用图结构来表示篇章,并参考 Hobbs 模型建立了篇章图库(Discourse GraphBank)①.篇章图库的标注来源是美联社新闻专线和华尔街日报,一共标注了 135 篇文章的图结构.

(3)借鉴 RST 标注的汉语篇章语料库

乐明[53-54]以 RST 理论为指导,同时参考了汉语复句和句群理论,进行了汉语篇章结构语料(财经评论,CJPL)标注的尝试,完成 97 篇人民网财经评论文章的修辞结构标注,标注了篇章单位、连接词、连接词的位置、修辞关系和连接词所在 EDU 的核心性等内容.

在该篇章语料库中,汉语修辞关系集有 12 组(例如并加关系组、选择关系组、对立关系组、条件关系组、原因关系组和背景关系组等)47 种汉语修辞关系,每个修辞关系都有后缀来区分该关系的篇章单位的核心性地位,例如"解答-N" 表示解答关系中起核心作用的篇章单位;"解答-S"表示解答关系中起卫星作用的篇章单位;"解答-M"表示解答关系中每个篇章单位都是核心成分,也表示该解答关系是一个多核关系.在该语料库中,乐明应用删除测试和替换测试来区分篇章单位的核心性地位,如果一个修辞关系的两个篇章单位难以判断谁是核心,就标注为多核结构.为检验独立标注的一致性,运用 SPSS 对 10 篇语料的修辞关系进行了一致性测算,其 Kappa 值为 0.638.

乐明的研究是汉语篇章语料库的建设上进行的有益尝试.该语料规模比较小,尚未有基于该语料的计算模型研究.

(4)基于连接依存树的汉语篇章树库

Li 等人[35]以连接依存树的篇章结构理论为基础,构建了汉语篇章结构语料库(Chinese Discourse TreeBank,CDTB).CDTB 采用自顶向下的标注策略,对每一段内容先找出其最上层关系,然后递归地对切分后的内容进行标注.目前 CDTB 共有 500 个文档,全部来自宾州汉语树库(Penn Chinese TreeBank,简称 CTB)[14],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 标注内容包括:连接词、连接词位置、连接类型、关系类型、中心位置、子句切分位置、子节点和父节点等.

① Wolf F, Gibson E, Fisher A, Knight M. Discourse Graphbank. Linguistic Data Consortium, Philadelphia, 2005. https://catalog.ldc.upenn.edu/LDC2005T08

CDTB 共标注篇章关系 7310 个,其中显式关系 1814 个,隐式关系 5496 个,篇章结构层次最大为 9 层. 篇章主次关系中,单中心关系 3555 个,多中心关系 3755 个. 单中心关系中,中心在前 2108 个,中心在后 1447 个.

RST-DT 和 CDTB 的结构相似,但因其语言和构建方法不同,存在一定的区别. 具体的区别有:(1)在基本篇章单位定义上,RST-DT 可以是短语或者子句,CDTB 的基本篇章单位一定有标点作为标志,一般是小于或等于句子的单位,即子句;(2)在连接词的处理上,RST-DT 没有考虑连接词;CDTB 参考 PDTB 对连接词的处理方式,将连接词作为谓词连接篇章单位;(3)在篇章关系的定义上,RST-DT 给出修辞关系的类别定义;CDTB 则将关系和连接词区分开,给出了一个通用的关系分类;(4)在结构表示上,CDTB 和 RST-DT 均可构建完整的篇章结构树;(5)在篇章单位主次区分上,RST-DT 在修辞关系定义中明确限定核心和卫星的判断标准,CDTB 根据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并不相同,例如"之所以……是因为……"和"因为……所以……"都是因果关系,但是"之所以……是因为……"的更强调原因项,而"因为……所以……"的更强调结果项. RST-DT 和 CDTB 更直观的对比情况汇总如表 1 所示.

**表 1 RST-DT 与 CDTB 的对比**

| 类别 | RST-DT | CDTB |
|---|---|---|
| 基本篇章单位 | 短语式子句,一个关系有一个或多个子句 | 子句,自顶向下切分,一个关系有两个或多个子句 |
| 连接词 | 没有考虑连接词 | 标识显式连接词和是否可删;添加隐式连接词 |
| 篇章关系 | 给定语义类别并进行标注 | 用连接词代表关系;将连接词映射到关系体系上;标注连接词及其属性 |
| 结构 | 可以构建完整篇章结构树 | 可以构建完整篇章结构树;自顶向下分割 |
| 主次关系 | 由修辞关系类别决定 | 由全局意图决定,与关系无直接关联 |

### 3.1.3 计算模型

在上述的语料资源上,针对篇章主次关系计算模型的研究主要集中在修辞结构篇章树库(RST-DT)和基于连接依存树的汉语篇章树库(CDTB)上,这是因为在这两个语料库上明确标注了微观的篇章主次关系.

(1)基于 RST-DT 的篇章主次关系计算模型

由于修辞结构理论明确定义了篇章微观主次关系("核心-卫星"结构,以及单核关系、多核关系),因此英文的篇章主次关系计算模型研究,主要集中在基于修辞结构篇章树库的篇章结构分析中. 然而,现有研究并没有将篇章的"主次关系"作为一个独立的研究对象,而是将其作为篇章修辞结构分析的一个辅助环节. 本章节将概要的介绍这些计算模型.

基于 RST-DT 的篇章结构分析主要包括两个子任务,分别是:基本篇章单位 EDUs 的划分和篇章结构的生成.

EDUs(Elementary Discourse Units,基本篇章单位)划分子任务主要是对篇章文本进行正确切割,相关研究较多. Soricut 和 Marcu[55] 采用概率模型结合最大似然估计和相应的数据平滑算法进行文本切分,他们的算法称为 SPADE,在标准句法树上获得了 84.7%的 $F1$ 性能. Tofiloski 等人[56] 利用句法和词汇等特征对进行 EDUs 的分割,在自动句法树上取得了 84%的 $F1$ 性能. Hernault 等人[57] 将 EDUs 的识别转化为序列标注问题,利用词汇、词性标注、中心词、句法等特征,采用 CRF 模型取得了 94%的 $F1$ 性能. 目前基于 RST-DT 的 EDUs 划分任务准确率已经比较高,提升空间不大.

而篇章结构生成方面,性能则不够理想. Soricut 和 Marcu[55] 在 SPADE 算法中,利用语法和词法信息进行句子级的篇章结构分析,其 $F1$ 性能达到 49.0%,主次关系识别 $F1$ 性能为 30.9%. Hernault 等人[57] 实现了基于 SVM 的篇章结构分析器 HILDA. 该分析器对篇章单位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,完整结构树识别的 $F1$ 值为 47.3%. LeThanh 等人[58] 在句子级别,采用句法信息和线索短语等信息生成篇章结构;在篇章级别,将相邻的句子信息和结构信息融入到集束搜索(Beam Search)算法中,从而降低了篇章结构生成的搜索空间,其 $F1$ 性能为 53.7%,主次关系识别 $F1$ 性能为 47.1%. Duverle 等人[59] 利用词汇、结构化成分和句法等特征进行篇章结构分析,取得了 48.1%的 $F1$ 性能. Feng 和 Hirst[60] 在 HILDA 的基础上,增加了语言学特征,其篇章结构生成的 $F1$ 性能为 55.87%. Joty 等人[61] 在他们前期句内篇章结构分析[62] 的工作基础上,分别应用句内和句间两个动态 CRF 模型,构建句子级别和篇章级别的分析器,使用句子级别的子树、滑动窗口和 CKY 算法自底向

上的构建篇章结构树，结构识别性能 $F1$ 为 55.71%. Ji 和 Eisenstein[63] 参考深度学习(Deep Learning)的做法，采用线性变换将表面特征转化成隐空间进行移进规约，使得主次关系识别率达到 71.13%.

在研究了 Soricut 等人、LeThanh 等人、Hernault 等人和 Joty 等人的篇章主次关系的研究工作后，能够发现篇章主次关系的识别与篇章结构生成具有强关联，他们的相关研究成果如表 2 所示. 这也充分说明了篇章主次关系识别之于篇章结构生成的重要性.

**表 2　基于 RST-DT 的篇章主次关系识别性能**

| 文献 | 篇章主次识别 | | | 篇章结构生成 |
|---|---|---|---|---|
| | 准确率/% | 召回率/% | $F$ 值/% | $F$ 值/% |
| Soricut 和 Marcu | 54.0 | 21.6 | 30.9 | 49.0 |
| LeThanh 等人 | 47.8 | 46.4 | 47.1 | 53.7 |
| Hernault 等人 | 61.9 | 58.3 | 60.0 | 47.3 |
| Joty 等人 | — | — | 68.4 | 55.7 |

(2) 基于 CDTB 的篇章主次关系计算模型

相对于英文篇章结构分析在 RST-DT(发布于 2002 年)上的相关工作，汉语篇章单位主次关系的分析研究开始的相对较晚. 其原因是，现有的汉语篇章语料资源构建工作都晚于英文的篇章语料资源，如前文 3.1.2 章节所述，乐明借鉴 RST 标注的汉语篇章语料库发布于 2008 年，Li 等人基于连接依存树的篇章树库发布于 2014 年，这两个语料库对微观的主次关系进行了标注，并且目前仅有 CDTB 上有相关的计算模型研究.

Li[34] 在 CDTB 语料库上构建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、隐式篇章关系识别和篇章单位主次识别等部件，最终输出一个完整的篇章结构树，其中子句识别在标准句法树上获得 95.1% 的 $F1$ 性能，连接词识别的 $F1$ 性能为 69.3%，隐式篇章关系识别的 $F1$ 性能因果类 32.4%、并列类 77.3%、解说类 51.8%，篇章结构识别在自动识别子句和自动句法树上的 $F1$ 性能为 46.4%. Chu 等人[64] 利用上下文特征、词与词性特征和词对特征等进行了汉语篇章主次关系识别的研究，针对中心在前、中心在后、多中心这 3 种类型的主次关系，识别性能 $F1$ 值分别为 51.58%、53.59%、54.64%.

### 3.2　篇章宏观主次关系

#### 3.2.1　理论体系

从宏观角度看篇章，篇章的整体结构与篇章体裁和篇章模式紧密相关，不同的体裁的展开形式是不同的，如新闻类篇章常用“总分”或“事件-评论”的结构来展开，而法院庭审的篇章核心框架则常常由“何方式、何因、何事、何据、何推断……”的结构构成.

图 6 是一个新闻体裁的图式结构(出自 van Dijk 在 1988 年的论著)[42]. 该结构中有两个主体部分：总述(Summary)和报道(Story)，总述包含标题(Headline)和导语(Lead)两个部分，而报道包括情景(Situation)与评论(Comments)两个部分，并通过一个从上而下的层级顺序展示了假拟的新闻篇章的宏观结构. 其中，标题和导语反映全文的中心话题，情景是对中心话题的详细描述(包括事件、结论、环境和背景等内容)，评论是记者对新闻事件的看法与评价.

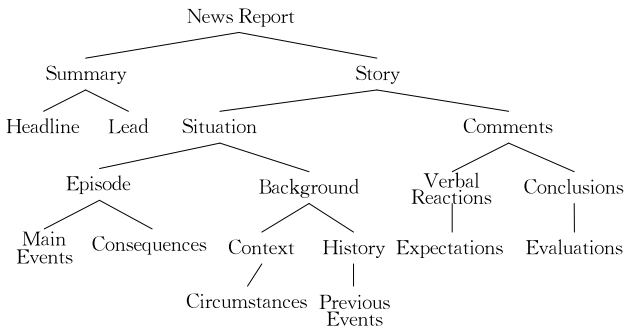
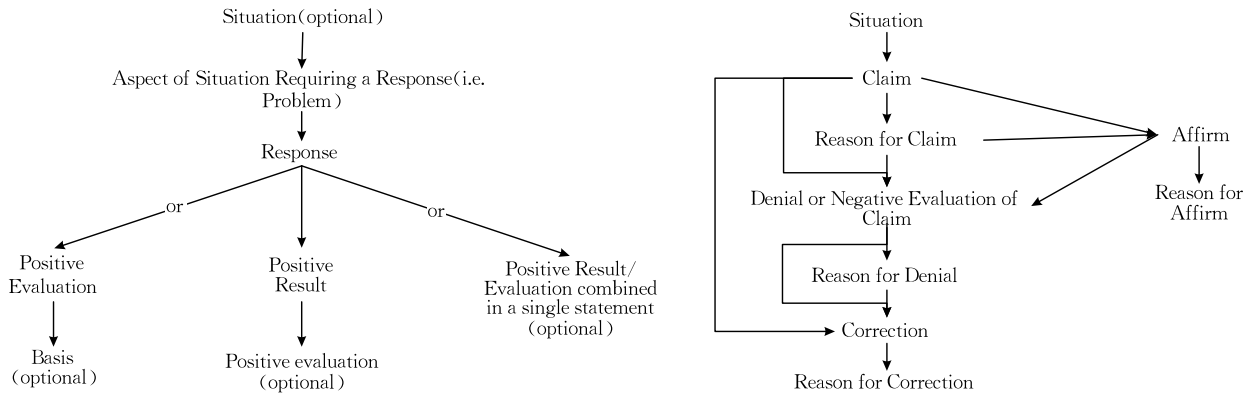



图 6　假拟新闻图式结构(出自 van Dijk 在 1988 的论著[42])

相同的体裁下也有不同的表达方式，即不同的篇章模式. 篇章模式是篇章组织的宏观结构，Hoey[65] 将篇章模式定义为“构成篇章关系的组合”. 篇章宏观模式与特定环境与篇章体裁有关，表达特定的意图. 经过反复使用，各种体裁形成了各自特定的、程式化的篇章组织结构和语言特征，这些结构模式和特征是同一领域的共同承认和遵守的规约. 根据语言学家的研究，常见的篇章模式有“问题-解决模式(Problem-Response Pattern)”、“提问-回答模式(Question-Answer Pattern)”、“主张-反主张模式(Claim-Counterclaim Pattern)”和“一般-特殊模式(General-Particular Pattern)”等[65-67]. 图 7(a)和图 7(b)分别表示了“问题-解决模式”和“主张-反主张模式”的篇章模式.

“问题-解决模式”一般包括 4 个部分，分别是“情景(Situation)”、“问题(Problem)”、“解决办法(Response)”和“评估(Evaluation)”，其中“问题”和“解决办法”是必备的部件，是篇章的核心，而其他部分，例如对现象或情景的描述、对解决办法的评价等都是可选部件，相对于必备部件来说，可选部件代表篇章中的次要信息. “主张-反主张模式”一般由“情景(Situation)”、“主张(Claim)”和“反主张(Denial)”




(a) 问题-解决模式 (b) 主张-反主张模式

图 7 问题-解决模式和主张-反主张模式

三大部分构成,其中“主张”和“反主张”是该模式的核心和必要成分,而其他部分都不是必备部件.从宏观角度来看,不同的篇章模式由不同的主要部分和次要部分构成,篇章主次关系的判别也需参考不同篇章模式的结构特点.

篇章宏观主次关系研究的相关理论主要包括 Martin 和 Rose 的超主位理论和 van Dijk 的宏观结构理论.

(1)超主位理论

Martin 和 Rose[37] 把 Halliday[28-29] 功能语法理论的“主位-述位”和“已知信息-新信息”概念映射到篇章分析层面,指出了篇章的每个段落都有一个“主题句(Topic sentence)”,从篇章的角度看,这个主题句可以被看作是段落层的主位,即“超主位(Hyper Theme)”,而超主位再向上还有“宏观主位(Macro Theme)”,它是更高层次的主位(例如篇章的标题),宏观主位之上还可以类推下去更高层次的宏观主位(例如书籍的目录).超主位和宏观主位,即宏观角度的主要信息,能帮助读者理解多篇章间的目录、篇章的主题和段落的主要内容.超主位后出现的是新信息,被称为“超新信息(Hyper New)”,超主位由其后的超新信息进一步解释、证明和发展,超新信息把该段落的内容推向一个新起点.比“超主位-超新信息”更高一层的还有“宏观主位-宏观新信息(Macro New)”.

Martin 和 Rose 把 Halliday 的“主位-新信息”结构放大到篇章层次上,形成波浪式的和有层级的“主位-新信息”、“超主位-超新信息”以及“宏观主位-宏观新信息”结构.

(2)宏观结构理论

篇章宏观结构理论由 van Dijk[38-42] 提出,该理论较为系统地描述了微观结构与宏观结构的层次关系:微观结构是篇章中的一个句子内部的结构或两个连续的句子之间的结构,表现的是篇章内部连续语句之间的语义连贯;而宏观结构是更高层次的结构,表现为篇章的主要思想、推进脉络和整体上的语义连贯,从宏观上约束了篇章内各部分之间的关联.

宏观结构理论指出,篇章的连贯性表现为两个层次:微观结构连贯和宏观结构连贯.微观结构连贯指线性或顺序性连贯,篇章中句子或一系列句子表达的命题意义之间相互联系构成一个连续的统一体.宏观结构连贯指总摄全篇的总主题所代表的语义结构由次级的主题共同蕴含,并层层分解,形成统一的整体.只有同时满足微观连贯和宏观连贯时,才能构成整个篇章的连贯.

篇章宏观结构具有相对性和层次性的特点.篇章各级的宏观结构都是由相对于较低层次的单元组合而来.因此篇章中存在着不同层次的宏观结构.从微观结构到宏观结构是一个循序渐进的推导过程.图 8 给出了篇章宏观结构分析的一个过程.

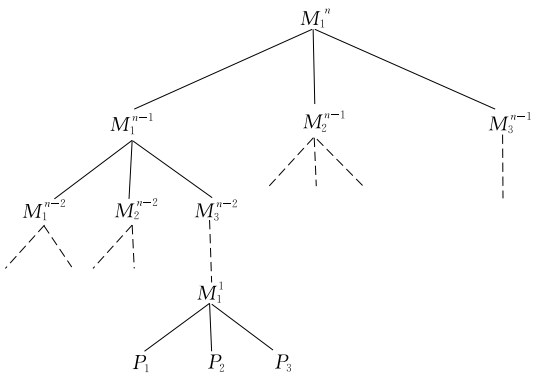



图 8 篇章宏观结构

在篇章宏观结构图中,$P_1$,$P_2$,$P_3$…表示较低层次的组合单元,而 $M_j^k$ 表示一个宏观单元,上标 $k$ 表示层次,下标 $j$ 表示宏观单元在当前层次的顺序.在

这个层次结构中，上层的单元涵盖下层单元的内容，上一层单元的结构和语义都是由它对应的下一层单元的共有命题组成的，依据宏观规则逐层推导各层篇章单元的主要信息，形成上一层篇章单元，直至最上层 $M^n$，形成一个多层次的篇章宏观结构. 如果在整个篇章的结构最上层有一个或一组命题能够包含整个篇章的命题，那么这整个篇章便是连贯的，最上层的命题表达的就是整个篇章的宏观语义结构.

van Dijk 提出了 4 条宏观规则来解决由微观结构向宏观结构推导的问题. (1) 删除规则. 删除两个或多个相关联篇章单位中不重要的信息；(2) 选择规则. 如果一个篇章单位($P_1$)的所表述的语义内容蕴含在另一个篇章单位($P_2$)中，那么选择 $P_2$ 而舍弃 $P_1$；(3) 概括规则. 用高一级的概括性抽象概念代替具体的表示部分特征的概念；(4) 归总规则. 把事件中属于同一个经验框架的信息用表述该经验框架的信息单元来覆盖. 其中删除规则和选择规则在实践中比较好操作，因为在篇章中存在被保留或选择的子句，仅需识别并选择主要的，舍弃次要的. 而概括原则和归总原则被概括或归总出来的子句在原篇章中并不存在，要求对篇章的语义有更深入的理解，才能获得较好的结果. 这 4 条宏观规则的形式化建模也是一个难点.

3.2.2　语料资源和计算模型

目前，尚未有相关文献涉及篇章宏观主次关系语料资源的构建和计算模型的研究.

## 4　存在问题和研究策略

通过对篇章主次关系研究现状的分析，可以发现，篇章主次关系的识别和分析，有利于理解篇章的中心主题、核心重点、展开思路和论据支撑作用等，也有利于挖掘篇章整体语义连贯及篇章内各部分之间的关联. 然而，尽管目前在微观和宏观的角度都有理论研究，但尚未有研究把微观和宏观的主次关系看作一个整体的研究对象. 事实上，从整个篇章的角度来看，微观主次关系和宏观主次关系相辅相成：在宏观主题和宏观结构的框架之下，篇章通过微观结构逐步推进和展开；而微观结构的形式和内容，也是为整个篇章的主旨和意图服务的. 因此将微观和宏观主次关系结合起来，构成统一的整体，更加符合篇章结构和语义分析的需求.

篇章主次关系的研究工作主要围绕理论体系、语料资源和计算模型这三个方面进行. 通过对已有研究的分析，我们总结出篇章主次关系研究中目前存在的问题，并提出相应的研究策略.

(1) 理论体系方面

尽管 1975 年在 Grimes 的研究中就已经提出了篇章衔接关系连接的单元存在主次之分，之后的研究中，微观角度有修辞结构理论，宏观角度有超主位理论和篇章宏观结构理论，但是尚没有统一的理论表示体系. 微观和宏观角度的篇章主次关系的侧重点不同：微观篇章主次结构的侧重点是句子内和句子间的修辞关系，不同的修辞关系，其对应的“核心”与“卫星”也不同；而宏观的主次关系侧重于研究段落间、章节间的主要和次要关系. 因此，微观和宏观主次关系的层次和判别方法都有一定的差异.

微观和宏观主次关系的特点不同，也应采用不同的策略. 微观层面，句子间的联系比较紧密，两个或多个篇章单位之间的关系，可以借鉴修辞结构理论和基于连接依存树的篇章结构理论进行分析和判断；宏观层面，段落间的逻辑关系不像句子间那么紧密，相对松散，侧重于段落间、章节间的关系以及段落与篇章主题的关联，可以结合篇章宏观结构等理论，并融入不同体裁文章的不同篇章模式，进行篇章结构分析和主次关系识别研究.

在判定篇章主次关系时，应对篇章的主题和主要内容有充分的认识，把握局部原则和全局原则. 局部原则是指判定的篇章单位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；全局原则指判定的篇章单位在符合局部原则的同时，需要结合全文的主旨来判定.

针对微观和宏观主次关系的特点，借鉴篇章宏观结构理论、修辞结构理论等，可以归纳出一个微观和宏观统一的篇章主次关系的结构化表示体系. 在该体系中，用篇章结构树的形式来表示篇章的层次关系，自顶向下地构建一个由篇章的标题、章节、段落、微观结构、篇章单位和基本篇章单位等构成的多级篇章结构. 其中，第一层为篇章标题层，第二层为章节层(可根据语料的实际情况而设定)，第三层为段落层，段落以下的层次为微观结构层. 上层结构与下层结构之间的连线表示其层次之间的整体与部分间关系；同级结构之间的连线表示其依存关系，并利用箭头的方向表示各级篇章结构间的主要和次要关系. 图 9 表示了一个微观和宏观统一的多层篇章主次关系结构.

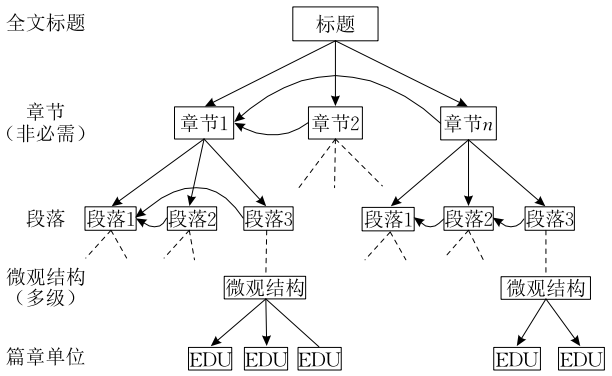



图 9 微观和宏观统一的篇章主次结构

在这个多层结构中，篇章标题层、章节层和段落层的主次结构属于宏观结构，而微观结构层相对复杂，可应用基于连接依存树的篇章结构理论来构建. 微观结构的连接依存树在前面章节中已经描述，这里不再赘述，参见 3.1.1 章节和图 1(例 1～3 的篇章结构连接依存树表示).

通过构建微观和宏观统一的篇章主次关系表示体系，能够更直观地理解篇章结构，进而更方便分析篇章内容. 在构建该篇章结构树的具体分析方法上，引入微观篇章结构分析方法和扩展修辞结构分析方法；在宏观结构分析中引入和扩展宏观结构理论分析方法；宏观结构分析能够指导微观结构的识别，而微观结构中的关键词和线索词等信息也将辅助宏观结构的生成.

(2)语料资源方面

英语篇章语料库修辞结构篇章树库(RST-DT)标注了篇章微观主次关系，即篇章单位的“核心性(Nuclearity)”，借鉴 RST 标注的汉语篇章语料库(CJPL)和基于连接依存树的汉语篇章树库(CDTB)也分别对篇章微观主次关系进行了标注. 但尚没有语料资源对篇章宏观主次关系进行标注和分析.

鉴于此，构建一个微观和宏观统一的篇章语料库非常必要. 进行微观和宏观主次结构的标注研究，不仅能够为篇章主次关系的研究和分析奠定基础，并且能为基于主次关系的篇章结构分析研究提供强有力的支撑.

我们已进行了篇章主次关系语料库的初步标注工作，根据微观和宏观统一的篇章主次关系表示体系，从宾州汉语树库(Penn Chinese TreeBank，CTB)中抽取部分文章作为语料来源. 在宏观角度，研究篇章主题、段落话题、章节间关系和段落间关系的标注方法、策略和结构. 微观角度与 CDTB 的微观主次关系标注结合起来，研究子句间关系和微观结构，进而建立一个微观与宏观统一的篇章主次关系标注体系. 语料建设采用自顶向下的标注策略，从宏观上把握篇章整体结构，以宏观结构的标注约束微观结构标注.

我们初步标注了 97 个篇章的宏观篇章结构(选取 CTB 语料中前 100 篇，去掉段落数为 1，不能形成段落间关系的 3 篇)，共标注了 533 个段落之间 433 个关系(其中多元关系都转换为二元关系保存)，平均段落数为 5.49 段/篇，最大段落数为 13 段. 表 3 为标注语料资源的统计数据.

在宏观篇章关系定义上，我们借鉴 CDTB 定义的篇章关系，根据语料的实际情况，整理选用调整了原有的关系定义，形成了三大类(并列类、因果类和解说类)15 种(如并列关系、顺承关系、原因-结果关系、解说关系和评价关系等)宏观的篇章关系，如图 10 所示.

**表 3 初步标注语料统计数据**

| 文档总数 | 段落总数 | 平均段落数(段/篇) | 最大段落数 | 最小段落数 | 句子总数 | 平均段落长度(句/段) | 段落关系总数(多元关系转换为二元关系) |
|---|---|---|---|---|---|---|---|
| 97 | 533 | 5.49 | 13 | 2 | 1339 | 2.51 | 433 |

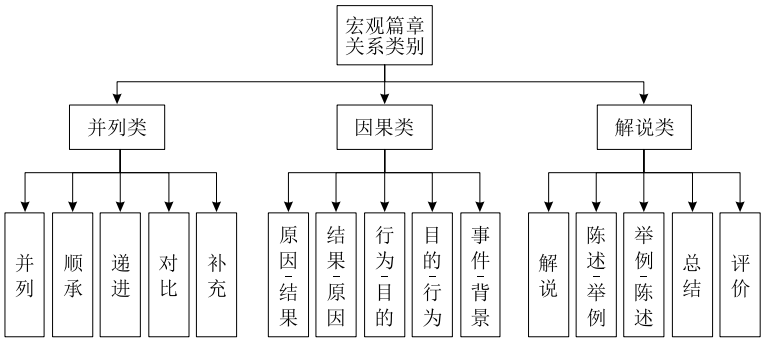



图 10 宏观篇章关系类别

已完成的标注工作显示：(1)树型结构是篇章结构较好的形式化表示方法，不仅能够有效的表示篇章的层次关系，还能直观表达篇章主次关系；(2)由于篇章关系类型较多，对主次关系的判断也存在多种可能，因此当前的标注工作仍是相当主观的，标注时需要对整个篇章有比较深刻的理解和认识；(3)篇章结构、篇章关系和篇章主次关系的标注之间具有很强的关联性和制约性，例如背景关系中，背景的重要性一定低于事件本身的重要性，在主次结构中处于次要地位；(4)篇章的篇幅越大，对篇章的理解难度就会越大，标注的一致性也相应的越低.

(3)计算模型方面

英文篇章主次关系的分析方法主要集中在基于RST-DT的篇章结构分析中，其中主次关系的识别作为篇章结构分析的一个重要组成部分，目前性能最好的是Ji和Eisentein的篇章结构分析模型，其主次关系识别性能为71.13%.相对于英文篇章分析，汉语篇章分析更具挑战.因为从语句衔接方式上来说，英语多用“形合法(Hypotaxis)”，而汉语多用“意合法(Parataxis)”，也就是说英语多在句法形式上使用连接性词语将句子或子句衔接起来，而汉语不一定依赖连接性词语而更多的是依靠意义上的衔接.汉语篇章主次关系分析的计算模型研究较为缺乏，Chu等人基于CDTB语料库进行了初步的汉语篇章主次关系研究，其识别率低于55%.上述计算模型都是针对篇章微观主次关系的识别，采用的研究方法也比较单一，主要利用词汇、句法等特征，利用最大熵和支持向量机等进行分类训练，没有利用更为有效的语义信息.

相对而言，由于篇章宏观主次关系的复杂性，目前的研究还停留在理论层面的探索，尚未有相应的计算模型.因此深入研究宏观主次关系的分析方法非常必要.深入探索从微观结构到宏观结构的语义转换规则，在van Dijk提出的四条宏观规则的基础上，考虑将宏观规则转化成形式化的表示方法和计算模型，用以从微观结构层次的推导出更大的篇章单位，直至推导出篇章宏观结构和篇章主题.

根据本文前期研究，主次关系的识别，对于篇章结构和内容的理解有重要的意义，很大程度上影响了篇章结构分析的性能.因此，在微观主次关系和宏观主次关系计算模型的研究基础上，探索微观和宏观主次关系之间的交互作用，构建一个微观主次关系和宏观主次关系联合学习的计算模型，将有利于充分利用微观和宏观的特点，提高篇章主次关系识别性能，进而提高篇章结构、篇章语义的分析性能.

## 5　未来研究方向

从目前国内外已有的研究成果和存在问题来看，篇章结构分析已经成为自然语言处理领域的研究热点和重点，而篇章主次关系的研究尽管在理论体系、语料资源、计算模型方面都还不够完善，但其重要性已经逐步显现出来，引起了研究者的重视.基于此，本文总结归纳了篇章主次关系未来的一些研究方向，供读者参考.

(1)篇章主次关系的基础研究

目前篇章主次关系的基础研究还不够完善，特别是宏观角度的研究更加匮乏，因此构建一套针对篇章主次关系的完整研究体系，包括理论体系、语料资源、计算模型，是非常必要的.

微观角度，由于句子间的联系比较紧密，侧重于篇章单位之间的关系，可以借鉴修辞结构理论和基于连接依存树的篇章结构理论进行篇章结构构建和主次关系分析；宏观角度，段落间的逻辑关系不像句子间那么紧密，而是侧重于段落间、章节间的关系以及与篇章主题的关联，可结合篇章宏观结构理论(例如，将van Dijk的宏观规则进行系统化、形式化建模)、主述位结构理论、信息结构理论等，并融入不同体裁文章的篇章模式，进行篇章宏观结构分析和段落主次关系的研究.结合微观和宏观主次关系的特点，以宏观结构的分析指导微观结构的识别，以微观结构中的关键词、线索词等信息辅助宏观结构的生成，形成统一的整体，发挥各自的优势，可相互促进微观和宏观主次关系分析性能的提高.

(2)基于篇章主次关系的篇章结构分析

在篇章主次关系分析的基础上，结合已有的篇章关系识别和篇章结构生成等篇章结构分析的研究成果，探索这些任务之间的交互作用和制约关系，构建联合学习的计算模型，可进一步提高篇章结构分析的整体性能.

篇章结构分析可分为子句识别、连接词识别和分类、隐式篇章关系识别、篇章主次关系识别以及篇章结构树构建等子任务.其中子句识别和连接词识别这两个子任务的性能已经比较好，达到90%以上，而隐式篇章关系识别、篇章主次关系识别和篇章结构树构这三个子任务的性能却在50%～60%之间，因此当前篇章结构分析的重点和难点是隐式篇章关系识别、篇章主次关系识别和篇章结构树构建.

鉴于这三个子任务之间存在相互的促进作用,可采用联合学习的方法将这三个子任务结合起来,找出子任务之间的相互约束关系,降低错误传递和子任务之间的相互影响.并且,单独进行各个子任务的工作时容易忽略一些跨子任务的全局性的特征,通过对全局参数的调节,利用联合学习进行任务的整合,将有利于篇章结构分析获得更好的性能.

(3)基于篇章主次关系的篇章语义分析

语义分析的目的是通过建立有效的模型和系统,将自然语言转化为形式语言,使计算机能够理解和使用自然语言,能与人类进行无障碍的沟通.语义分析分为词汇级、句子级和篇章级三个层次,目前词汇级和句子级的语义分析都已取得一些研究成果,并在统计机器翻译、自动问答和信息抽取等领域中得到了应用,而篇章级语义分析还停留在初始阶段.

篇章级语义分析不仅需要理解篇章中信息的组织结构方式,获取篇章中包含的信息,还需要在词汇分析和信息抽取的基础上分析篇章中句子段落之间的内在逻辑结构关系,并以此来构建整个篇章的组成架构,最终理解整个篇章表达的语义信息.

篇章是一个多平面,多层级的语言体系,词汇、句子、段落和章节之间,互相制约、互相作用和互相协调,取得语义的一致性,才能具有较好的衔接性和连贯性,也就是说词、句和章之间的联系构成了整个篇章的语义结构.

词汇系统内部词汇义位之间的语义关系表现为聚合关系,而篇章中的词汇义位之间的语义关系除了聚合关系以外,还表现为组合关系.聚合关系在词典中已经确定了语义关系,相对固定;而组合关系很多情况是特定语境中表现出来的语义关联,是一种非固定的关系.因此需要在上下文环境中寻找词汇义位之间的语义关联,并形成由中心词汇构成的核心语义场.在同一篇章内,这些词汇通过义位的相互渗透和相互制约,取得语义的一致性,词汇和句子之间的关联方式不同,构成了不同的篇章语义.

篇章级语义分析的目的就是通过篇章结构的理解,更好的理解篇章的语义,而主次关系的识别,对于篇章结构和语义的理解具有重要的意义.从语义分析的角度,更重要的语义信息来源于篇章的主要部分,而非次要部分,最能代表篇章主要思想的内容来源于最主要的篇章单位.

因此篇章主次关系的研究,将通过更准确的构建篇章结构,把语义分析的重点定位在主要篇章单位上,从篇章主题和宏观结构出发把握篇章重点,并根据上下文的词汇义位分析语义关联,再结合词汇级和句子级分析的研究方法,这或将给篇章级的语义分析带来新的进展.

(4)基于篇章主次关系的应用研究

篇章是一个有机的结构体,结构分析清楚了,文章的核心部分自然更方便找到,因此,可以充分考虑上下文信息,利用微观和宏观统一的主次关系语料,对整个文档集进行建模,构建一个层次化主题模型.基于这个层次化主题模型,可进一步进行篇章主题抽取和自动文摘的应用研究,以期提高篇章应用的效率和准确率.

目前对文档进行主题模型建模的常用方法包括 TF-IDF 方法、潜在语义分析(Latent Semantic Analysis,LSA)[68-69]、概率浅层语义索引(probabilistic Latent Semantic Indexing,pLSI)[70]和浅层狄利赫雷分配(Latent Dirichlet Allocation,LDA)[71]等.传统 LDA 模型只能在词汇层进行主题分析.因此,可考虑在扩充 LDA 模型的基础上,利用标注的微观和宏观统一的篇章主次关系语料,对整个文档集进行建模,形成一个层次化主题模型,该模型将包括篇章单位、段落、文档和文档集合等层次.图 11 表示了该层次化主题模型.

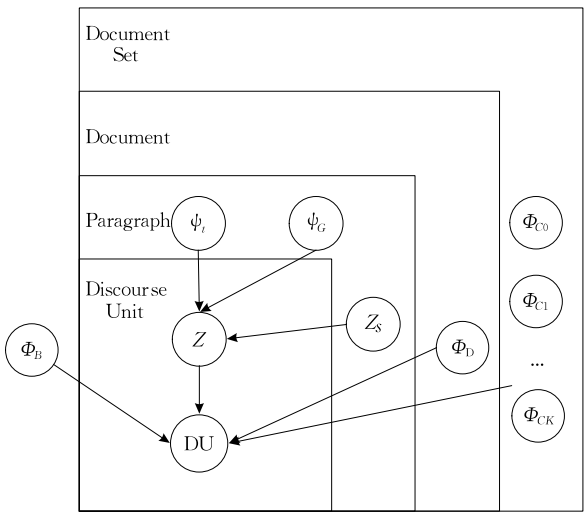



图 11 层次化主题模型的图形化表示

通过这个模型,可以分别体现文档集合与文档、文档与段落以及段落与篇章单位之间的主题关系,揭示不同文本单元主题之间的层次关系,并有效利用段落间的关联强度、篇章主次结构等信息.这个模型将直接有助于主题抽取应用研究.

自动文摘根据文摘选取方式不同可分为抽取型文摘和理解型文摘.由于理解型文摘对语言生成技术要求较高,目前的方法难以付诸实用,因此抽取型

文摘是主要的研究方向[72-74]，采用的方法以基于统计的方法居多.由于基于统计的方法本质上是考虑了句子的词汇特征(例如关键短语、单词和短语的词频等)，没有考虑句子间结构关系和文档的主题分布，忽略了文档与文档之间的信息，因此自动文摘的系统性能出现了瓶颈，生成的摘要信息冗余大、主题分布不均衡.

将层次化主题模型应用到自动文摘的应用研究中，运用句子间结构关系、文档间主题关系等信息，结合已有的以基于词汇的统计方法，通过定位篇章中最重要的章节和段落，并以这些章节和段落的主要信息为自动文摘的信息来源，以期提高自动文摘应用的效率和准确率.

## 6　小　结

篇章主次关系是篇章结构分析的重要环节.我们研究发现，篇章主次关系的研究与分析，有助于理解篇章的结构和语义，并为自然语言处理的应用提供有力的支持.然而，已有的研究一般只是将篇章主次关系识别看作篇章结构生成的一个辅助环节，篇章主次关系研究不仅未受到足够重视，并且成为了篇章结构分析的一个瓶颈.因此，本文将篇章主次关系从篇章结构识别中分离出来，作为一个独立的研究对象进行系统和深入的研究.

本文针对篇章主次关系这个研究内容，探讨了什么是篇章主次关系以及为什么要对篇章主次关系进行研究，然后从微观和宏观两个角度，详细阐述了篇章主次关系在理论体系、语料资源和计算模型等三个方面的研究现状，总结了目前存在的问题并提出了相应的研究策略和初步方案，最后展望了篇章主次关系未来的研究方向.

理论体系的研究和建设是自然语言处理领域研究的基石，理论体系的完整性和可操作性将直接影响到语料资源的质量和计算模型的性能，最终影响到自然语言处理的应用性能.语料资源是基于统计的自然语言处理研究的基础资源，构建一个高质量、大规模的语料库具有较高的研究和应用价值.因此，针对篇章主次关系的研究重点是构建一个微观和宏观统一的理论体系，并建设相应的汉语篇章主次关系语料资源，打好坚实的研究基础.在这个基础上，进一步深入研究篇章主次关系对篇章结构分析和篇章语义理解的影响，从而提高篇章分析的整体性能.篇章分析的最终目的还是为自然语言相关应用提供服务，因此可将篇章主次关系的研究成果应用到主题抽取、自动文摘和问答系统等应用研究中，以期得到更高性能.

## 参　考　文　献


[1] de Beaugrande R, Dressler W. Introduction to Text Linguistics. London, UK: Longman, 1981

[2] Meyer T, Popescu-Belis A. Using sense-labeled discourse connectives for statistical machine translation//Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation. Jeju Island, South Korea, 2012: 129-138

[3] Guzmán F, Joty S, Màrquez L., Nakov P. Using discourse structure improves machine translation evaluation// Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 687-698

[4] Peldszus A, Stede M. Joint prediction in MST-style discourse parsing for argumentation mining//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 938-948

[5] Atkinson J, Munoz R. Rhetorics-based multi-document summarization. Expert Systems with Applications, 2013, 40(11): 4346-4352

[6] Ferreira R, de Souza Cabral L, Freitas F, et al. A multi-document summarization system based on statistics and linguistic treatment. Expert Systems with Applications, 2014, 41(13): 5780-5787

[7] Cohan A, Goharian N. Scientific article summarization using citation-context and article's discourse structure//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 390-400

[8] Liakata M, Dobnik S, Saha S, et al. A discourse-driven content model for summarizing scientific articles evaluated in a complex question answering task//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle, USA, 2013: 747-757

[9] Presutti V, Draicchio F, Gangemi A. Knowledge extraction based on discourse representation theory and linguistic frames//Proceedings of International Conference on Knowledge Engineering and Knowledge Management. Galway, Ireland, 2012: 114-129

[10] Zou B, Zhou G, Zhu Q. Negation focus identification with contextual discourse information//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014: 522-530

[11] Mukherjee S, Bhattacharyya P. Sentiment analysis in twitter with lightweight discourse analysis//Proceedings of the 24th International Conference on Computational Linguistics. Mumbai, India, 2012: 1847-1864

[12] Mittal N, Agarwal B, Chouhan G, Bania N, Pareek P. Sentiment analysis of hindi review based on negation and discourse relation//Proceedings of the 6th International Joint Conference on Natural Language Processing. Nagoya, Japan, 2013: 45-50

[13] Bhatia P, Ji Y, Eisenstein J. Better document-level sentiment analysis from RST discourse parsing//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 2212-2218

[14] Xue N. Annotating discourse connectives in the Chinese Treebank//Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, USA, 2005: 84-91

[15] Halliday M A K. Linguistic function and literary style: An inquiry into the language of William Golding's the Inheritors// Chatman S ed. Literary Style: A Symposium. London: Oxford University Press, 1971(Vol. 339)

[16] Hobbs J R. Information, intention, and structure in discourse// Proceedings of the NATO Workshop on Burning Issues in Discourse. Istanbul, Turkey, 1993: 41-66

[17] Hobbs J R. Coherence and coreference. Cognitive Science, 1979, 3(1): 67-90

[18] Mann W C, Thompson S A. Relational propositions in discourse. Discourse Processing, 1986, 9(1): 57-90

[19] Mann W C, Thompson S A. Rhetorical structure theory: A theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse, 1987, 8(3): 243-281

[20] Mann W C, Matthiessen C, Thompson S A. Rhetorical structure theory and text analysis. Discourse Description: Diverse Linguistic Analysis of a Fund-Raising Text, 1992: 39-78

[21] Marcu D. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts[Ph. D. dissertation]. Department of Computer Science, University of Toronto, Toronto, Canada, 1997

[22] Marcu D. The Theory and Practice of Discourse Parsing and Summarization. Cambridge, MA: MIT Press, 2000

[23] Prasad R, Dinesh N, Lee A, et al. The penn discourse TreeBank 2.0//Proceedings of the Language Resources and Evaluation Conference. Marrakech, Morocco, 2008: 2961-2968

[24] Zhou Y, Xue N. PDTB-style discourse annotation of Chinese text//Proceedings of the Association for Computational Linguistics. Jeju Island, South Korea, 2012: 69-77

[25] Zhou Y, Xue N. The Chinese discourse TreeBank: A Chinese corpus annotated with discourse relations. Language Resources and Evaluation, 2015, 49(2): 397-431

[26] Grosz B J, Sidner C L. Attention, intentions, and the structure of discourse. Computational Linguistics, 1986, 12(3): 175-204

[27] Grosz B J, Weinstein S, Joshi A. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics, 1995, 21(2): 203-225

[28] Halliday M A K. An Introduction to Functional Grammar. London, UK: Arnold, 1985

[29] Halliday M A K. An Introduction to Functional Grammar (2nd Edition). London, UK: Arnold, 1994

[30] Webber B. D-LTAG: Extending lexicalized TAG to discourse. Cognitive Science, 2004, 28(5): 751-779

[31] Wu Wei-Zhang, Tian Xiao-Lin. The Chinese Sentence Group. Beijing: The Commercial Press, 2000(in Chinese)
(吴为章, 田小琳. 汉语句群. 北京: 商务印书馆, 2000)

[32] Xing Fu-Yi. Research on Chinese Complex Sentence. Beijing: The Commercial Press, 2001(in Chinese)
(邢福义. 汉语复句研究. 北京: 商务印书馆, 2001)

[33] Yao Shuang-Yun. A Research on the Collocation of the Relation Markers of Chinese Compound Sentences and Some Relevant Explanation[Ph. D. dissertation]. Central China Normal University, Wuhan, 2006(in Chinese)
(姚双云. 复句关系标记的搭配研究及相关解释[博士学位论文]. 华中师范大学, 武汉, 2006)

[34] Li Yan-Cui. Research of Chinese Discourse Structure Representation and Resource Construction[Ph. D. dissertation]. Soochow University, Suzhou, 2015(in Chinese)
(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]. 苏州大学, 苏州, 2013)

[35] Li Y, Feng W, Sun J, Kong F, Zhou G. Building Chinese discourse corpus with connective-driven dependency tree structure//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 2105-2114

[36] Hoey M. On the Surface of Discourse. Buckley, Washington, USA: George Aller, and Unwin Publisher, Ltd., 1983

[37] Martin J R, Rose D. Working with Discourse: Meaning Beyond the Clause. London: Continuum, 2003

[38] van Dijk T A. Macrostructure: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, Inc, Publishers, 1980

[39] van Dijk T A. Text and Context: Explorations in the Semantics and Pragmatics of Discourse. London, UK: Longman, 1977

[40] van Dijk T A, Kintsch W. Strategies of Discourse Comprehension. New York, USA: Academic Press, 1983

[41] van Dijk T A. Handbook of Discourse Analysis. 4 Vols. London, USA: Academic Press, 1985

[42] van Dijk T A. News as Discourse. Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, Inc, Publishers, 1988

[43] Grimes J. The Thread of Discourse. The Hague: Mouton, 1975

[44] Zhu Q, Li J, Wang H, Zhou G. A unified framework for scope learning via simplified shallow semantic parsing// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Massachusetts, USA, 2010: 714-724

[45] Lin Z, Kan M Y, Ng H T. Recognizing implicit discourse relations in the Penn Discourse TreeBank//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 343-351

[46] Kong F, Zhou G, Zhu Q. Employing the centering theory in pronoun resolution from the semantic perspective//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore, 2009: 987-996

[47] Xu F, Zhu Q, Zhou G. A unified framework for discourse argument identification via shallow semantic parsing//Proceedings of the 24th International Conference on Computational Linguistics. Mumbai, India, 2012: 1331-1340

[48] Kong F, Ng H T, Zhou G. A constituent-based approach to argument labeling with joint inference in discourse parsing//Proceedings of the Empirical Methods in Natural Language Processing. Doha, Qatar, 2014: 68-77

[49] Xu Fan. Zhu Qiao-Ming, Zhou Guo-Dong. Implicit discourse relation recognition based on tree kernel. Journal of Software, 2013, 24(5): 1022-1035(in Chinese)
(徐凡，朱巧明，周国栋. 基于树核的隐式篇章关系识别. 软件学报，2013，24(5)：1022-1035)

[50] Xu Fan. Research of Key Issues in English Discourse Structure Analysis [Ph.D. dissertation]. Soochow University, Suzhou, 2013(in Chinese)
(徐凡. 英文篇章结构分析关键问题研究[博士学位论文]. 苏州大学，苏州，2013)

[51] Carlson L, Marcu D, Okurowski M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory//van Kuppevelt J, Smith R W eds. Current and New Directions in Discourse and Dialogue. Springer Netherlands, 2003: 85-112

[52] Wolf F, Gibson E. Representing discourse coherence: A corpus-based study. Computational Linguistics, 2005, 31(2): 249-287

[53] Yue Ming. Annotation and Analysis of Chinese Financial News Commentaries in Terms of Rhetorical Structure[Ph. D. dissertation]. Communication University of China, Beijing, 2006(in Chinese)
(乐明. 汉语财经评论的修辞结构标注及篇章研究[博士学位论文]. 中国传媒大学，北京，2006)

[54] Yue Ming. Rhetorical Structure Annotation of Chinese News Commentaries. Journal of Chinese Information Processing, 2008, 22(4): 19-23(in Chinese)
(乐明. 汉语篇章修辞结构的标注研究. 中文信息学报，2008，22(4)：19-23)

[55] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—Volume 1. Edmonton, Canada, 2003: 149-156

[56] Tofiloski M, Brooke J, Taboada M. A syntactic and lexical-based discourse segmenter//Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing. Suntec, Singapore. 2009: 77-80

[57] Hernault H, Prendinger H, Ishizuka M. HILDA: A discourse parser using support vector machine classification. Dialogue & Discourse, 2010, 1(3): 1-33

[58] LeThanh H, Abeysinghe G, Huyck C. Generating discourse structures for written texts//Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland. 2004: 329-335

[59] Duverle D A, Prendinger H. A novel discourse parser based on support vector machine classification//Proceedings of Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing: Volume 2—Volume 2. Suntec, Singapore, 2009: 665-673

[60] Feng V W, Hirst G. Text-level discourse parsing with rich linguistic features//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers—Volume 1. Baltimore, USA, 2012: 60-68

[61] Joty S, Carenini G, Ng R T Mehdad Y. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis//Proceedings of the 51st Annual Meeting of Association for Computational Linguistics. Sofia, Bulgaria, 2013(1): 486-496

[62] Joty S, Carenini G, Ng R T. A novel discriminative framework for sentence-level discourse analysis//Proceedings of the Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, South Korea, 2012: 904-915

[63] Ji Y, Eisenstein J. Representation learning for text-level discourse parsing//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore, USA, 2014 (1): 13-24

[64] Chu X, Wang Z, Zhu Q, Zhou G. Recognizing nuclearity between Chinese discourse units//Proceedings of the 19th International Conference on Asian Language Processing. Suzhou, China, 2015: 197-200

[65] Hoey M. Textual Interaction: An Introduction to Written Discourse Analysis. London: Routledge, 2000

[66] Huang Guo-Wen. Theory and Practice of Discourse Analysis. Shanghai: Shanghai Foreign Language Education Press, 2001 (in Chinese)
(黄国文. 语篇分析的理论与实践. 上海：上海外语教育出版社，2001)

[67] Hu Shu-Zhong. Studies in English Text Linguistics. Shanghai: Shanghai Foreign Language Education Press, 2005(in Chinese)
(胡曙中. 英语语篇语言学研究. 上海：上海外语教育出版社，2005)

[68] Landauer T K, Foltz P W, Laham D. An introduction to latent semantic analysis. Discourse Processes, 1998, 25(2-3): 259-284

[69] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 1990, 41(6): 391-407

[70] Hofmann T. Probabilistic latent semantic indexing//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Berkeley, USA, 1999: 50-57

[71] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation. The Journal of Machine Learning Research, 2003, 3(1): 993-1022

[72] Louis A, Nenkova A. A coherence model based on syntactic patterns//Proceedings of the Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, South Korea, 2012: 1157-1168

[73] Almeida M B, Martins A F T. Fast and robust compressive summarization with dual decomposition and multi-task learning //Proceedings of the 51st Annual Meeting of Association for Computational Linguistics. Sofia, Bulgaria, 2013(1): 196-206

[74] Li C, Qian X, Liu Y. Using supervised bigram-based ILP for extractive summarization//Proceedings of the 51st Annual Meeting of Association for Computational Linguistics. Sofia, Bulgaria, 2013(1): 1004-1013

**CHU Xiao-Min**, born in 1981, Ph.D. candidate. Her research interests include natural language processing and discourse analysis.

**ZHU Qiao-Ming**, born in 1963, Ph.D., professor, Ph. D. supervisor. His research interests include natural language processing and Web information processing.

**ZHOU Guo-Dong**, born in 1966, Ph. D., professor. His research interests include natural language processing, information extraction, statistical machine translation and machine learning.

**Background**

Discourse primary-secondary relationship analysis aims to recognize the primary and secondary parts in discourse structures. It is an important research topic in natural language processing due to its critical role in document-level discourse structure analysis. However, recent studies in discourse analysis totally ignore the importance of recognizing primary-secondary relationship, and view it as a dispensable component attached to the analysis of rhetorical structure.

Generally speaking, there exist two hierarchical levels of primary-secondary relationship: micro level and macro level. Most popular discourse-related theories are developed at the micro level, such as rhetorical structure theory, discourse purpose theory, and connective-drive dependency tree theory, etc. As a result, there exist several corpora annotated under the guide of those theories. On the contrast, very few theories are proposed at the macro level.

In this paper, we explore to recognize the primary-secondary relationship which plays a critical role in discourse structure analysis. Specifically, in contrast to existing work, we view it as an independent task from the discourse rhetorical structure analysis. This paper summarizes the research status of discourse primary-secondary relationship analysis through discussing the key issues and challenges in discourse analysis, and presenting our resolutions as well as future work.

Discourse primary-secondary relationship analysis not only is helpful in understanding the discourse center themes, discourse structure and semantics, but also provides strong support for deep applications in natural language processing, such as summarization, topic extraction, question answering, etc.

In recent years, the authors' NLP group has been specially focusing on discourse structure analysis. The achievements include proposing connective-driven dependency tree theory, building a Chinese discourse structure corpus consisting of 500 documents from Penn Chinese TreeBank, and developing a Chinese discourse parser with state-of-the-art performance.

This work is supported by the National Natural Science Foundation of China (61272260), the Ministry of Education China Mobile Research Foundation (MCM20150602), and the Jiangsu Provincial Science and Technology Plan (BK20151222).