软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(2):321−343 [doi: 10.13328/j.cnki.jos.005868]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

宏观篇章结构表示体系和语料建设*

褚晓敏[1], 奚雪峰[2], 蒋 峰[1], 徐 昇[1], 朱巧明[1], 周国栋[1]

[1](苏州大学 自然语言处理实验室,江苏 苏州 215006)

[2](苏州科技大学 电子与信息工程学院,江苏 苏州 215009)

通讯作者: 朱巧明, E-mail: qmzhu@suda.edu.cn

摘 要: 篇章结构分析是自然语言处理领域的一个重要研究方向.篇章结构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(如自动文摘、信息抽取、问答系统等)提供有力的支撑.目前,篇章结构分析主要集中在微观的层面,分析的重点是句子内部或句子与句子之间的关系和结构,而宏观层面的研究相对较少.因此,以篇章结构作为研究对象,并将研究重点放在宏观篇章结构的表示体系和语料资源建设上.探讨了篇章结构分析的重要性,从理论体系、语料资源、计算模型这 3 个方面阐述了篇章结构分析的研究现状,提出了以篇章主次关系为媒介的宏观和微观统一的篇章结构表示框架,并分别构建了宏观篇章的逻辑语义结构和功能语用结构.在此基础上,标注了规模为 720 篇新闻报道的宏观篇章结构语料,并对标注的结果进行了一致性分析和标注统计分析.

关键词: 篇章结构分析;宏观篇章结构;篇章结构表示体系;逻辑语义结构;功能语用结构;语料标注

中图法分类号: TP18

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/5868.htm

Macro Discourse Structure Representation Schema and Corpus Construction

CHU Xiao-Min[1], XI Xue-Feng[2], JIANG Feng[1], XU Sheng[1], ZHU Qiao-Ming[1], ZHOU Guo-Dong[1]

[1](Natural Language Processing Laboratory, Soochow University, Suzhou 215006, China)

[2](School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China)

Abstract: Discourse structure analysis is an important research topic in natural language processing. Discourse structure analysis not only helps to understand the discourse structure and semantics, but also provides strong support for deep applications of natural language processing, such as automatic summarization, information extraction, question answering, etc. At present, the analysis of discourse structure is mainly concentrated on the micro level. The analysis focuses on the relations and structures between sentences or sentences groups, while the analysis on macro level is less. Therefore, this study takes discourse structure as the research object, and focuses on the construction of representation schema and corpus resources on the macro level. This study discusses the importance of discourse structure analysis, expounds the research status of discourse structure analysis from three aspects, namely, theory system, corpora resource, and computing model, and puts forward the macro-micro unified discourse structure representation framework with the primary-secondary relation as the carrier. Furthermore, this study constructs the logical semantic structure and functional pragmatic structure of macro discourse level respectively. On this basis, this study annotates a macro Chinese discourse structure corpus, consisting of 720 newswire articles, and analyzes the results of the annotations in consistency and statistical data.

* 基金项目: 国家自然科学基金(61773276, 61673290, 61836007)
Foundation item: National Natural Science Foundation of China (61773276, 61673290, 61836007)
收稿时间: 2018-01-09; 修改时间: 2019-03-25, 2019-04-19; 采用时间: 2019-05-19; jos 在线出版时间: 2019-08-09
CNKI 网络优先出版: 2019-08-12 12:08:20, http://kns.cnki.net/kcms/detail/11.2560.TP.20190812.1208.011.html

Key words: discourse structure analysis; macro discourse structure; discourse structure representation schema; logical semantic structure; functional pragmatic structure; corpus annotating

自然语言处理是人工智能领域中的一个重要方向,其研究目的是实现人与计算机之间使用自然语言进行有效通信.自然语言处理的研究成果,让人类可以用日常的语言来使用计算机,而无需再花大量的时间和精力去学习计算机的语言.

篇章是由连续的话段或句子构成的语言整体,表达一个完整的语言信息.其特点是前后衔接、语义连贯,且具有一定的交际目的和功能.无论在形式上还是意义上,篇章都不是孤立存在的,而是由篇章单元各自承担一定的角色,相互作用,并通过篇章关系关联在一起,共同构成连贯的篇章结构,表达一定的篇章语义和意图.因此,我们解释、分析一个篇章的前提,就需要理解篇章的结构和语义.篇章结构分析是自然语言处理领域的一个重要研究方向,其研究目的是研究自然语言文本的内在结构,理解文本单元间的逻辑语义关联,即挖掘出文本的结构化和语义信息.篇章结构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(如自动文摘[1−3]、机器翻译[4−6]、信息抽取[7,8]、问答系统[9]等)提供有力的支撑.此外,篇章结构分析可以帮助计算机识别不同写作风格,并为文本自动生成奠定基础.

篇章结构的研究分析根据文本层次结构可分为微观和宏观两个层面:在微观层面,篇章结构指的是以句子为主体的篇章单元之间的结构与关系(主要包括子句与子句、句子与句子、句群与句群);在宏观层面,篇章结构指的是段落及以上的篇章单元之间的结构与关系(主要段落与段落、章节与章节、篇章与篇章).下面以两个具体的例子进一步阐明微观篇章结构和宏观篇章结构的含义和区别[10].

首先,以基于连接依存树的汉语篇章树库(Chinese discourse treebank,简称 CDTB)[11]中的一篇新闻报道 chtb_0035 中的第 2 段为例,说明微观篇章结构的含义.根据标注,每一个段落可以形成一棵完整的篇章结构树.

例 1:a. 〈例如〉记者了解到,去年天津对俄罗斯出口贸易额为两千五百万美元;b. 进口两千多万美元;c. 其中,出口额比九四年增长超过 80%;d. 此外,天津已有十多家企业在俄开设了公司或代表处;e. 俄罗斯在天津的投资企业也达 20 余家;f. 外方投资额约一千万美元;g. 双方经贸关系正稳步发展.

在本例中,叶子节点代表最小篇章单元(一般为子句),非叶子节点是连接词,连接词意味着篇章关系,箭头指向篇章关系中主要的篇章单元.在图 1 表示微观篇章结构树中,“〈·〉”括号内的连接词表示此处原本没有连接词,但可以手工添加,而 null 表示没有连接词并且不可添加.根据该结构图,可以帮助读者直观地理解该篇章的结构,另外,连接词可以帮助读者理解篇章单元间的篇章关系,即篇章的逻辑语义.

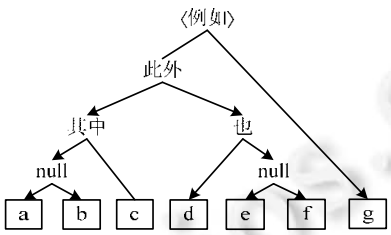

Fig.1 Micro discourse structure of Example 1

图 1 例 1 的微观篇章结构

我们再以宾州中文树库(the Penn Chinese treebank 8.0,简称 CTB 8.0)[12,13]中的一篇新闻报道 chtb_0094 为例,说明什么是宏观篇章结构.

例 2:标题:海南洋浦开发区将动工兴建一批工业项目.

- (P1). 新华社海口一月六日电(记者柳昌林)经过5年多的开发建设,海南洋浦经济开发区迎来工业建设高潮,5 个工业启动项目有的竣工投产,有的即将动工兴建.
- (P2). 洋浦经济开发区管理局局长王永春说,开发区招商工作取得突破性进展,开发区建设已由土地开

发迈向工业项目建设的新阶段.

- (P3). 日前,洋浦开发区工业启动项目之一的 60 万吨木浆厂已获国务院批准兴建.这个全国最大规模的木浆厂将由新加坡亚洲浆纸业股份有限公司投资 12.83 亿美元兴建,年产漂白商品木浆 60 万吨.
- (P4). 首个工业启动项目——金岛精米加工厂已于去年底竣工投产.这个由澳门远东(泰国)集团公司与海南省粮油集团公司等联合投资三千万美元兴建的精米加工厂,采用国外 90 年代最先进的生产设备和工艺流程,年加工 30 万吨糙米,加工后的精米 70%出口.
- (P5). 开发区的其他 3 个启动项目高速线材厂、橡木地板厂、浮法玻璃厂也将动工投产.此外,开发区已确定和可能确定的工业项目还有 20 多个,包括油汽化工、钢铁厂、还原铁等,总投资约 70 亿美元.
- (P6). 洋浦位于海南西部,是中国第一例由外商成片承包开发的工业开发区,享有目前国内最优惠、最开放的政策.过去 5 年间,土地开发商共投入 40 亿港元用于电厂、区内主干道及地下管网、土地平整、邮电通讯等基础设施建设.(完).

这篇新闻稿由 6 个段落 P1~P6 组成,段落与段落之间通过篇章关系连接在一起.在图 2 表示的宏观篇章结构树中,叶子节点表示段落,非叶子节点表示篇章关系,箭头指向主要篇章单元.根据该结构图,可以直观地理解该整个篇章的结构和篇章单元间的逻辑语义关系.

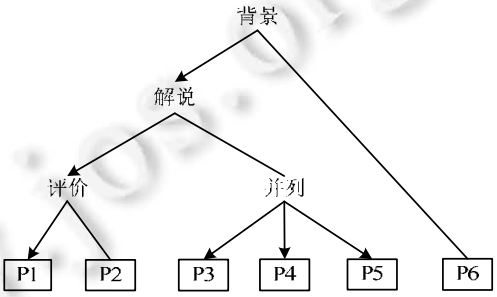

Fig.2 Macro discourse structure of Example 2

图 2 例 2 的宏观篇章结构

从微观和宏观这两个例子中可以看出,篇章结构的分析有助于篇章内容和主旨的理解.如例 1,EDU a~f 描述了天津与俄罗斯之间的经贸发展情况,是 EDU g 的实际例子,因此该段落的重点是 EDU g“双方经贸关系正稳步发展”,利用此篇章结构可以直观地理解段落的重点,进而理解作者的写作意图.如例 2,段落 P1 点明全文的主旨“海南洋浦经济开发区迎来工业建设高潮,5 个工业启动项目有的竣工投产,有的即将动工兴建.”而其后的段落都为这个全文主旨服务:P2 评价了这个事件的意义,P3~P5 展示“5 个工业启动项目”的具体进展情况,P6 介绍了整个“海南洋浦经济开发区”的发展状况.因此,整个篇章的重点是 P1,既是全文的主要内容,也是文章的主旨.

根据这样的分析可以预见,基于篇章结构分析可以进一步提升自然语言处理上层应用的性能,例如,应用篇章结构的信息总结篇章内容和篇章主旨、利用篇章关系的信息辅助问答系统的构建、利用主次关系提升自动文摘应用的性能等.然而,现有的篇章结构研究主要集中在微观层面,并且性能也未达到可以应用的水平.而宏观层面的研究还停留在理论研究上,尚没有可用的语料资源,也未见相应的计算模型.基于以上所述的原因,本文提出了一种宏观与微观统一的多层篇章结构表示框架,并分别构建了宏观篇章的逻辑语义结构和功能语用结构.在此基础上,本文标注了规模为 720 篇新闻报道的宏观汉语篇章语料库(原始语料来源为宾州中文树库 CTB 8.0).

本文的主要贡献如下.

(1) 提出一种宏观与微观统一的多层篇章结构表示框架,以主次关系为媒介将篇章宏观结构和微观结构统一起来,形成一个整体.

(2) 分别构建了宏观篇章结构表示体系的两个视图:逻辑语义视图和功能语用视图.基于这两个视图,可以有效地表达篇章结构、篇章关系、主次关系和篇章单元的角色功能,有助于整个篇章结构和语义

的理解.

(3) 根据上述宏观篇章结构表示体系,标注了 720 篇新闻报道的宏观篇章结构,其中包含 3 985 个段落、2 870 个篇章关系、8 319 个句子,共计 398 829 个字,为宏观篇章结构计算模型研究奠定坚实的基础.

本文第 1 节对篇章结构分析的相关工作进行了总结,阐述了篇章结构分析在理论体系、语料资源和计算模型等 3 个方面的研究进展.第 2 节提出一种以篇章主次关系为媒介的宏观与微观统一的多层篇章结构表示框架.第 3 节和第 4 节分别阐述了宏观篇章结构表示体系的逻辑语义视图和功能语用视图.第 5 节对语料标注的策略、方法、过程等进行了详细的描述.第 6 节对标注后的语料进行了统计和分析.第 7 节对全文进行总结,并对未来的研究内容进行初步探讨.

1 相关工作

篇章结构的研究分析,根据文本层次结构可分为微观篇章结构和宏观篇章结构:在微观层面,篇章结构指的是以句子为主体的篇章单元之间的结构与关系(主要包括子句与子句、句子与句子、句群与句群);在宏观层面,篇章结构指的是段落及以上的篇章单元之间的结构与关系(主要段落与段落、章节与章节、篇章与篇章).

1.1 理论体系

微观篇章结构理论包括浅层衔接理论[14]、Hobbs 模型[15,16]、修辞结构理论[17−19]、宾州篇章树库理论[20−22]、意图结构理论[23,24]、句群理论[25]、复句理论[26,27]、基于连接依存树的篇章结构理论[11,28]等.宏观篇章结构理论相对较少,包括篇章模式[29]、超主位理论[30]、篇章宏观结构理论[31−35]等.

(1) 修辞结构理论

修辞结构理论(rhetorical structure theory,简称 RST)创立于 20 世纪 80 年代,是 Mann & Thompson[17−19]在系统功能理论框架下提出的篇章生成和分析的理论.修辞结构理论的提出,对篇章结构分析的发展有着非常重要的意义,在这一理论框架下,研究人员进行了诸多的研究工作,标注了语料资源,构建了计算模型,并取得了越来越高的系统性能.修辞结构理论认为篇章各小句之间并非是杂乱无章的,而是通过修辞关系关联在一起的.修辞结构理论最初定义了 23 种篇章关系,每个修辞关系可以连接两个或多个篇章单元,修辞关系连接的两个或多个篇章单元中可能存在主要的篇章单元和次要的篇章单元,主要篇章单元与次要篇章单元之间形成“核心-卫星(nuclear-satellite)”的结构,其中“卫星”单元辅助“核心”单元的表达.当两个以上的篇章单元通过修辞关系连接在一起时,就构成修辞结构树.

(2) 基于连接依存树的篇章结构理论

Li 等人[11]借鉴修辞结构理论的树型结构表示,参考宾州篇章树库理论对连接词的处理方式,提出了基于连接依存树的篇章结构理论(connective-driven dependency tree,简称 CDT),进行了完整的篇章结构定义和描述.定义的内容包括基本篇章单元、连接词、篇章结构、篇章关系、篇章主次等.在该理论的连接依存树结构中,叶子节点表示基本篇章单元(elementary discourse unit,简称 EDU),内部节点为连接词(connective),箭头指向主要篇章单元.各层篇章单元通过连接词形成更高一级的篇章单元,层层组合,最终形成一棵完整的篇章结构树,如图 1 所示.

(3) 宏观结构理论

van Dijk[31−35]于 1980 年提出的宏观结构理论(macrostructure theory)指出:微观结构是一个句子内部或两个连续的句子之间的结构,表现语句之间的语义连贯;而宏观结构是更高层次的结构,表现为篇章整体上的语义连贯.宏观结构理论强调了篇章需要有一个总摄全篇的主题,并层层分解,由下层命题展开,最终形成统一的整体.微观和宏观的连贯是一体的,语句之间的微观连贯性为篇章整体的宏观连贯性服务.

1.2 语料资源

现有的篇章语料资源中,英文的主要包括宾州篇章树库(PDTB) 、修辞结构理论篇章树库(RST-DT)等,中文的主要包括基于连接依存树的汉语篇章树库(CDTB)和借鉴 RST 标注的汉语篇章语料库(CJPL)等.

(1) 修辞结构理论篇章树库(RST-DT)

修辞结构理论篇章树库(rhetorical structure theory-discourse treebank,简称 RST-DT)[36]由美国南加利福尼亚大学和华盛顿国防部以修辞结构理论为基础标注,并通过 Linguistic Data Consortium(LDC)于 2003 年发布,RST-DT 共标注了 385 个来自《华尔街日报》的篇章,标注了 53 种单核关系和 25 种多核关系,总词数 176 000 个.RST-DT 标注的篇章可以构成完整的篇章结构树.

(2) 宾州篇章树库(PDTB)

宾州篇章树库(Penn discourse treebank,简称 PDTB)[37]是由宾夕法尼亚大学、意大利托里诺大学和英国爱丁堡大学联合标注的,分为 3 个层次:第 1 层包括 4 类语义,第 2 层包括 16 类语义,第 3 层包括 23 类语义.PDTB 为每一种关系和其论元标注属性信息,以《华尔街日报》为语料来源,共标注了 2 304 篇文章,约 100 万字.PDTB 标注的篇章以论元对为标注对象,所标注的篇章不能构成完整的篇章结构树.

(3) 借鉴 RST 标注的汉语篇章语料库(CJPL)

乐明[38]以修辞结构理论为指导,参考汉语复句和句群的理论,标注了篇章单元、连接词、修辞关系、核心性等信息.共完成了 97 篇人民网财经评论的文章的篇章结构标注,标注了 12 组 47 种修辞关系.

(4) 基于连接依存树的汉语篇章树库(CDTB)

Li 等人[11]以基于连接依存树的篇章结构理论为基础,构建了包含子句切分位置、连接词信息、篇章关系、主次关系等信息的汉语篇章语料库(CDTB).目前,CDTB 共有 500 个文档组成,包含 2 342 个段落,每个段落形成一棵独立的篇章结构树,共标注了 7 310 个关系,标注了 4 大类 17 种篇章关系.

1.3 计算模型

在上述的语料资源上,可以形成完整的篇章结构是修辞结构篇章树库(RST-DT)和基于连接依存树的汉语篇章树库(CDTB)上,因此,本节重点介绍基于这两个语料库的计算模型研究.

(1) 基于 RST-DT 的计算模型

基于 RST-DT 的篇章结构分析包括两个子任务,分别是基本篇章单位 EDU 的划分和篇章结构的生成.

- 基本篇章单位 EDU 划分子任务的目标是将篇章文本切割为基本篇章单元.

Soricut 等人[39]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(SPADE).该算法在标准句法树上获得了 84.7%的 *F*1 性能.Hernault 等人[40]将 EDU 划分问题转化为序列标注问题,采用 CRF 模型,利用词汇、词性、中心词、句法等语言学特征取得了 94%的 *F*1 性能.目前,该任务已经达到了较高的性能.

- 相对而言,篇章结构生成任务的性能还有较大的提升空间.

Hernault 等人[40]实现的篇章结构分析器(HILDA)采用贪婪的自底向上的方法构建篇章结构树,使用 SVM 训练了篇章关系分类器,HILDA 篇章结构识别的 *F*1 性能为 72.3%、主次关系识别的 *F*1 性能为 59.1%、篇章关系识别 *F*1 性能为 47.8%、全树识别的 *F*1 性能为 47.3%.Joty 等人[41]在前期仅针对句内构建篇章结构分析器[42]的工作基础上,采用动态 CRF 模型,分别构建了句子级和篇章级两个篇章结构分析器,篇章结构识别的 *F*1 性能为 82.74%、主次关系识别的 *F*1 性能为 68.40%、篇章关系识别 *F*1 性能为 55.71%.Feng 和 Hirst[43]在 HILDA 的基础上增加了语言学特征,使用了线性链的条件随机场,在进行后编辑方法使得篇章结构识别任务达到了 85.7%的正确率、主次关系识别的正确率达到 71.0%、篇章关系识别的正确率达到 58.2%.Ji 等人[44]参考深度学习的做法,先采用线性变换将表面特征转化成隐空间,再进行移进规约构建篇章结构树,篇章结构识别的 *F*1 性能达到 86.0%、主次关系识别的 *F*1 性能达到 72.4%、篇章关系识别 *F*1 性能达到 59.7%.

专注在宏观篇章层面的研究较少,Sporlederd 等人[45]对 RST-DT 修正和裁剪后进行了宏观篇章结构分析,只达到了 59.0%的 *F*1 性能.

(2) 基于 CDTB 的计算模型

相对于英文篇章结构分析在 RST-DT 上的研究,汉语篇章结构分析开始得较晚.主要原因是,现有的汉语篇章语料资源的构建和发布都晚于英文的篇章语料资源.如前文所述,RST-DT 发布于 2003 年,乐明标注的 CJPL 发布于 2008 年,Li 等人[11]标注的 CDTB 发布于 2014 年,且目前仅有基于 CDTB 的汉语篇章结构计算模型研究.

李艳翠[28]在 CDTB 上构建了汉语篇章结构分析平台,分别进行了子句识别、关系识别、主次关系识别等子任务的研究分析,最终将子任务融合,输出完整的汉语篇章结构树.系统整体的性能(结构+主次+关系)在自动识别子句和自动句法树上仅达到 20.0%.Chu 等人[46]在 CDTB 上利用上下文特征、词与词性特征、词对特征等进行了篇章主次关系识别的研究,宏观 *F*1 值为 53.3%.

2 宏观与微观统一的多层篇章结构

本文针对宏观篇章结构目前语料资源和计算模型的研究都不充分的问题进行了深入研究,提出利用篇章主次关系作为媒介,将宏观和微观两个层面的篇章结构连接在一起.并以宏观篇章结构为主要研究对象构建了表示体系和语料资源,为后续的宏观篇章结构分析奠定基础.

针对微观和宏观篇章结构的特点不同,应采用不同的分析策略:微观层面,由于句子间的联系比较紧密,篇章结构分析依赖于篇章单元之间的逻辑语义因此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结构分析和关系判断;宏观层面,段落间的逻辑关系则相对松散,一个篇章单元本身即可以表达完整的语义,宏观篇章结构分析更依赖于篇章单元在全文中的地位和作用(如篇章单元与篇章主题的关联),因此,可以结合篇章宏观结构理论、超主位理论,进行篇章结构分析研究.

通过对现有研究内容的分析,我们发现,现有理论体系尽管定义不同,并且主要集中在微观层面,但无论篇章关系和篇章结构如何定义,篇章的主次关系的判断方法和表现形式却非常类似.因此,本文针对这一特点,借鉴篇章宏观结构理论、修辞结构理论、基于连接依存树的篇章结构理论等理论框架以篇章主次关系作为媒介将篇章的宏观结构和微观结构统一起来,归纳出一个统一的多层篇章结构表示框架.

在该框架中,用一个树型结构来表示篇章的层次关系,自顶向下地包含篇章的标题、章节、段落、微观结构、最小篇章单元等多级篇章结构元素.其中,第 1 层为篇章标题层,第 2 层为章节层,第 3 层为段落层,段落以下的层次为微观结构层,而微观结构又由篇章单元和最小篇章单元构成.在上述层次结构中,段落层和微观结构依据逻辑语义形成多层结构.篇章单元与篇章单元之间通过篇章关系联系起来,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.这个多级篇章结构框架如图 3 所示.

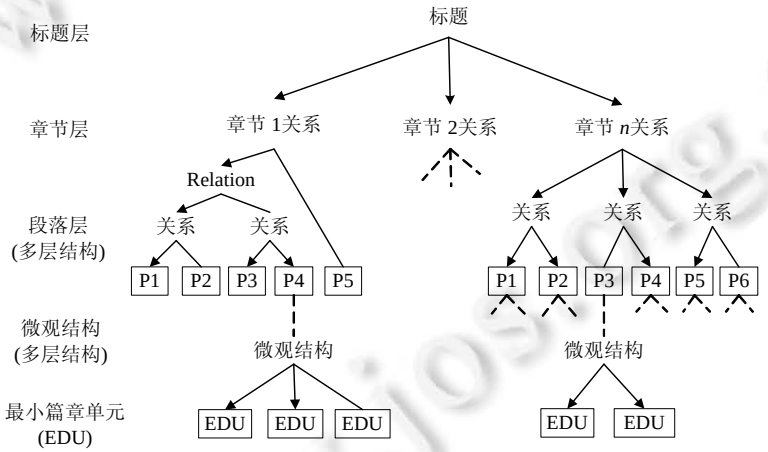

Fig.3 Macro and micro unified discourse structure representation framework

图 3 宏观和微观统一的篇章结构表示框架

(1) 标题层(title layer):标题层的内容即为文章的标题,代表全文的主题和主旨.

(2) 章节层(chapter layer):章节层的内容是文章分割出的章节.例如学术论文,其包括“引言、相关工作、系统模型、实验验证”等章节;又如一些杂志的文章可能分割成多个子标题.根据实际文章情况,章节层不是一个必需存在的层次.

(3) 段落层(paragraph layer):本文以自然分割的段落为该层次的对象,即段落层由自然分割的段落集合组

成.段落与段落之间由宏观篇章关系连接在一起,表达特定的逻辑语义.

(4) 微观结构(microstructure):一个自然段落形成一个微观结构.微观结构内包含多层篇章单元(discourse unit,简称 DU).篇章单元是指两个或两个以上的基本篇章单元(EDU)通过篇章关系联系在一起所形成的更大的单元结构.

(5) 基本篇章单元:篇章中最小的语义单位,具有独立性,表达一个基本的、完整的语义.由标点分割,至少包含一个谓语部分,表达一个命题.

该多层篇章结构的表示以篇章主次关系为媒介,尽管微观和宏观语义和表达方式有比较大的差异,但整个篇章的主次关系是一体的.即宏观篇章结构和语义需要微观篇章结构和语义的支撑,宏观篇章的连贯性需要微观篇章的连贯性来表达.只有在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.

本文的关注点是篇章的宏观结构,即段落层以上的篇章结构,因此,本文构建了一套宏观篇章结构(macro discourse treebank,简称 MDT)表示体系,用于表示宏观篇章结构.而微观结构部分,本文复用了 Li 等人[10]提出的基于连接依存树的篇章结构理论定义,在文中不再赘述,可参考相应的文献.

本文提出的宏观篇章结构表示体系将篇章结构用篇章结构树的形式来表示,并分别定义了逻辑语义和功能语用两种宏观结构视图,以及叶子节点、非叶子节点、边指向等结构要素.逻辑语义结构的侧重点是篇章单元与篇章单元之间的逻辑关系.功能语用结构的侧重点是篇章单元的功能,即在全文中的所承担的角色和所起的作用.本文第 3 节和第 4 节分别阐述了 MDT 的两个视图:逻辑语义结构和功能语用结构.

3 宏观篇章逻辑语义结构

逻辑语义结构的侧重点是篇章单元与篇章单元之间的逻辑关系.篇章单元是通过篇章关系连接在一起的,研究者通过分析篇章单元的语义和篇章单元之间存在的逻辑联系来明确篇章关系.

3.1 叶子节点

不同于微观结构将子句或谓语短语定义为叶子节点,在宏观篇章结构中,我们关注的是段落以上的层次之间的关系.因此,我们直接将自然分割的各个段落作为叶子节点.自然分割的段落,即作者在写作时根据其写作意图和逻辑语义所分割的段落,以换行符区分.这些自然分割的段落,也是宏观篇章结构中最小的篇章单元.

例如,在例 2 的篇章 chtb_0094 中,《海南洋浦开发区将动工兴建一批工业项目》中有 6 个段落,那么这 6 个段落 P1~P6,就是宏观篇章结构的叶子节点,这些段落形成的逻辑语义结构如图 2 所示.

3.2 非叶子节点和篇章关系

各层的篇章单元都是通过篇章关系连接在一起的,一个篇章关系连接两个或多个篇章单元,因此,我们将篇章关系视为宏观篇章结构的非叶子节点.

本文篇章关系的定义主要参考修辞结构理论,并结合汉语表达的习惯和宏观篇章结构的特点做整理和调整.主要调整的原因是:针对于宏观篇章的最小篇章单位为段落,而修辞结构理论的篇章关系划分是子句适用的,在颗粒度上有差异,造成一些篇章关系因为划分过细而无法适用的情况.表 1 对比了修辞结构理论和宏观篇章表示体系中的篇章关系定义.具体说明如下.

(1) 在宏观篇章结构中,最小的篇章单元是段落,在一个段落中常常既描述事件发生的时代背景,又阐明了事件发生所在的地理环境、周边情况等要素,而极少会出现分开描述这些信息的情况,因此,本文将环境关系(circumstance)和背景关系(background)合并为背景关系(background),以覆盖所有关于环境和背景的描述.

(2) 在汉语篇章中,常用“摆事实、讲道理、列数据”等手段来解释说明文章的主题,在一个篇章单元中可能包含多种解释说明的手段,因此,这一类篇章关系不宜划分过细,本文将解答关系(solutionhood)、解说关系(elaboration)、证明关系(justify)和解释关系(interpretation)合并为解说关系(elaboration).

(3) 在修辞结构理论中对于因果类关系划分很细,并区分意愿性和非意愿性,这对于段落为最小单元的宏

观结构来说难以准确区分,因为一个段落阐述的原因或结果可能既包含意愿性的,也包含非意愿性,因此,本文将意愿性原因关系(volitional cause)和非意愿性原因关系(non-volitional cause)合并为果因关系(result-cause),将意愿性结果关系(volitional result)和非意愿性结果关系(non-volitional result)合并为因果关系(cause-result).

(4) 使能关系(enablement)、动机关系(motivation)和目的关系(purpose)也常常会在一起描述,并且非常抽象,难以区分,因此,本文将这 3 种关系合并为行为目的关系(behavior-purpose).

(5) 重述关系(restatement)重点是重述前文内容,总结关系(summary)重点是将前文的内容做概括总结,这在宏观篇章中几乎都在同一个段落中出现,因此,本文将这两种关系合并为总结关系(summary).

(6) 评价关系(evaluation)、序列关系(sequence)、对比关系(contrast)和证据关系(evidence)在宏观篇章中都适用,本文保留并沿用了这些关系.

(7) 对照关系(antithesis)、让步关系(concession)、条件关系(condition)和析取关系(otherwise)这 4 种关系一般只出现在句内和句间,在宏观结构上并不适用,因此本文没有纳入这些关系.

(8) 还有一些篇章关系在修辞结构理论最初定义的时候没有被包含,本文增加了 Purpose-Behavior(目的行为关系)用于表达目的在前行为在后的篇章关系,递进关系(progression)用于表示后一个篇章单元在前一个篇章单元的基础上更进一步的篇章关系,补充关系(supplement)用于表示后一个篇章单元对前一个篇章单元进行额外信息补充的篇章关系,举例陈述关系(illustration-statement)用于表示举例在前、陈述观点在后的篇章关系.

Table 1 Comparison of discourse relations between RST and MDT

表 1 修辞结构理论和宏观篇章结构的篇章关系对比

修辞结构理论	宏观篇章结构
环境关系(circumstance) 背景关系(background)	背景关系(background)
解答关系(solutionhood) 解说关系(elaboration) 证明关系(justify) 解释关系(interpretation)	解说关系(elaboration)
意愿性原因关系(volitional cause) 非意愿性原因关系(non-volitional cause)	果因关系(result-cause)
意愿性结果关系(volitional result) 非意愿性结果关系(non-volitional result)	因果关系(cause-result)
使能关系(enablement) 动机关系(motivation) 目的关系(purpose)	行为目的关系(behavior-purpose)
重述关系(restatement) 总结关系(summary)	总结关系(summary)
评价关系(evaluation)	评价关系(evaluation)
序列关系(sequence)	顺承关系(sequence)
对比关系(contrast)	对比关系(contrast)
连接关系(joint)	并列关系(joint)
证据关系(evidence)	陈述举例关系(statement-illustration)
对照关系(antithesis) 让步关系(concession) 条件关系(condition) 析取关系(otherwise)	在宏观结构中不会出现
基础定义中无	目的行为关系(purpose-behavior) 递进关系(progression) 补充关系(supplement) 举例陈述关系(illustration-statement)

依据上文对于篇章关系的分析,本文将宏观篇章结构整理归纳为 3 大类(并列类、因果类、解说类)15 种篇章关系,见表 2.参考修辞结构理论和基于连接依存树的篇章结构理论,我们的篇章关系也是一个可扩展的集合,

可根据需要增加新的篇章关系.

Table 2 Macro discourse relations

表 2 宏观篇章关系

类别	篇章关系
并列类(coordination)	并列关系(joint)、顺承关系(sequence)、递进关系(progression)、对比关系(contrast)、补充关系(supplement)
因果类(causality)	因果关系(cause-result)、果因关系(result-cause)、背景关系(background)、行为目的关系(behavior-purpose)、目的行为关系(purpose-behavior)
解说类(elaboration)	解说关系(elaboration)、总结关系(summary)、评价关系(evaluation)、陈述举例关系(statement-illustration)、举例陈述关系(illustration-statement)

篇章关系的具体定义如下.

(1) 并列类(coordination)

- 并列关系(joint):多个篇章单元并列叙述相关的几件事情或同一事件/事物的几个方面,在语义上它们是并存、平行的关系.并列的篇章单元之间不分主次,逻辑上不相互依赖.一般情况下,可任意交换其顺序,而对全文的语义表达没有影响.
- 顺承关系(sequence):多个篇章单元之间存在时间或者步骤上的先后顺序,在语义上存在先后相承的关系.这些篇章单元不可随意交换顺序.
- 递进关系(progression):后一个篇章单元的意义比前一个篇章单元更进了一步,程度更深或意义更大.
- 对比关系(contrast):两个篇章单元描述的是对立的看法或态度,或者是事物的两个对立的方面.如“机会”与“挑战”.
- 补充关系(supplement):一个篇章单元对另一篇章单元添加了一部分的说明和解释,是对已有内容的添补.补充篇章单元对前述篇章单元起到了内容上的补充作用,一般来说,即使去除补充单元也不影响全文内容和语义的表述.

(2) 因果类(causality)

- 因果关系、果因关系:原因和结果是揭示客观世界中普遍联系着的事物具有先后相继、逻辑相连的一对范畴.原因是指引起一定现象的现象,结果是指由于原因的作用而引起的现象.两个篇章单元一个描述原因,一个描述结果,形成因果关系或果因关系.原因在前、结果在后,称为因果关系(cause-result);结果在前、原因在后,称为果因关系(result-cause).
- 行为目的关系、目的行为关系:一个篇章单元描述的内容是为了产生另一篇章单元描述的事件或事物而执行的行为和动作.行为在前、目的在后,称为行为目的关系(behavior-purpose);目的在前、行为在后,称为目的行为关系(purpose-behavior).
- 背景关系(background):一个篇章单元描述的内容,是对另一篇章单元描述的内容起作用的历史情况或现实环境.如时代背景、政治背景、历史背景、地理环境等.

(3) 解说类(elaboration)

- 解说关系(elaboration):解说的篇章单元是对被解说的事物、事件进行进一步的解释说明,或对事物、事件的几个方面进行详细的解释说明.解说的方法有概括解说、定义解说、分类解说、数字解说、引用解说等.常见情况之一,解说关系连接的两个篇章单元呈现“总分”结构,即一个篇章单元为总述单元,另一个篇章单元(该篇章单元常常由一组并列的篇章单元组合而成)为分述单元,分述单元从不同方面或不同角度对总述单元做了进一步的细化.
- 陈述举例关系、举例陈述关系:一个篇章单元通过列举实际的事例来阐述另一个篇章单元描述的事物、事件或观点,进一步增加事物、事件或观点的可信度.陈述在前、举例在后,称为陈述举例关系(statement-illustration);举例在前、陈述在后,称为举例陈述关系(illustration-statement).
- 总结关系(summary):后一个篇章单元是对前一个篇章单元的客观重述或概括总结.在结构上,总结呈现

出先分述后总述的“分总”结构,这是与解说关系最大的区别;在内容上,总结不带有主观色彩,仅仅是对另一个篇章单元的内容做概括,这是与评价关系最大的区别.

- 评价关系(evaluation):后一个篇章单元是对前一个篇章单元描述内容的判断、分析或看法,常常含有主观色彩.

举例说明,在图 2 所示的篇章结构图中,“评价”“并列”“解说”和“背景”是非叶子节点.P2 评价了开发区招商工作,与 P1 构成评价关系(evaluation).P3~P5 分别介绍了“60 万吨木浆厂”“金岛精米加工厂”和“其他 3 个启动项目”的情况,用来详细说明 P1 中提到的“海南洋浦开发区”的“5 个工业启动项目”,这 3 个段落之间形成并列关系(joint).P3~P5 形成的整体与 P1,P2 形成的整体构成解说关系(elaboration).P6 介绍了洋浦的基本情况以及过去 5 年的基础设施建设,是全文的背景,因此与 P1~P5 形成的整体构成背景关系(background).

3.3 边指向和篇章主次关系

一个篇章关系连接两个或多个篇章单元,这些篇章单元隶属于同一个关系组,如果其中一个篇章单元可以概括该关系组的意图和内容,并且可以代表所在的关系组和外界发生关系,那么这个篇章单元就是一个主要单元,而其他的篇章单元则是次要单元.如果两个或两个以上的篇章单元都是同等重要的,那么这个篇章关系连接的篇章单元就没有主次之分.

根据这样的现象,我们定义了 3 种类型的篇章主次关系.分别是:

- 主要-次要关系(primary-secondary,简称 PS),这种关系中,主要篇章单元在前,次要篇章单元在后.
- 次要-主要关系(secondary-primary,简称 SP),这种关系中,次要篇章单元在前,主要篇章单元在后.
- 同等重要关系(equal importance,简称 EI),这种关系中,篇章单元同等重要,不分主次.

例如,在“举例陈述关系(statement-illustration)”中,前一个篇章单元负责陈述信息,而后一个篇章单元举例说明前一个陈述的篇章单元,使得陈述单元更容易被读者所理解,所以“举例陈述关系”是一个 PS 关系;再例如,“并列关系(joint)”中两个或多个篇章单元之间没有主次之分,是同等重要的,因此“并列关系”是一个 EI 关系.

在宏观篇章结构表示体系中,我们利用边的指向来区分主要的篇章单元和次要的篇章单元,带箭头的边指向主要篇章单元,而不带箭头的边指向次要篇章单元.举例说明:在图 2 所示的篇章结构图中,P1 指出洋浦经济开发区的“5 个工业启动项目有的竣工投产,有的即将动工兴建”,比 P2 对该事件的评价与主题更为相关,因此 P1 是主要的,P2 是次要的,形成 PS 关系.P3~P5 形成的并列关系,分别介绍这 5 个工业启动项目,它们之间没有主次之分,同等重要,因此构成 EI 关系.P3~P5 详细说明 P1 所提到的内容,该解说关系前面是主要的,后面是次要的,形成 PS 关系.而 P6 作为背景的介绍,其内容是次要的,而 P1~P5 所述内容是主要的,因此形成 PS 关系.依据图中的箭头指向,可以直观地表示篇章的主次关系,定位全文的主要内容.

3.4 完整的篇章结构树

最后,本文将前述章节定义的叶子节点、非叶子节点、边指向相结合,自上而下的分析篇章结构,并生成完整的篇章结构树.在这个逻辑语义结构树中,以自然分割的段落作为叶子节点,篇章关系作为非叶子节点,箭头指向篇章关系中主要的篇章单元.采用树型结构来表示宏观篇章结构是因为树型结构的形式清晰直接,也符合一般的中文阅读习惯;同时,树结构也能够直观地表示篇章的层次和篇章单元间的主次关系.

4 宏观篇章功能语用结构

功能语用结构的侧重点是篇章单元的功能,即在全文中所承担的角色和所起的作用.每一个篇章单元都不是无用的,而是承担了一定的功能,为全文的内容和主旨服务.van Dijk[34]在 1988 年指出,新闻篇章具有形象、具体的“图式结构(news schemata)”.该结构主要包括总述(summary)和故事(story),总述又包括标题(headline)和导语(lead)两部分,故事又包括情节(situation)和评论(comments).图式结构通过一个自顶向下的层级顺序展现整个新闻篇章的宏观形式,如图 4 所示.Fairclough[47]将图式结构中的情节和评论更名为附属部分(satellite)和收尾部分(wrap-up).Bell 等人[48]认为,新闻故事框架可由属性(attribute)、概述(abstract)和故事(story)这 3 大部分组成:

属性部分指明了新闻的基本要素(包括记者、事件发生的时间、地点等);概述部分包括标题及导语,概要地介绍新闻的主题内容;而故事部分由一个或多个情节(episode)构成,详细地描述新闻报道的细节.有一些新闻还包含3个补充成分,如背景(background)、评论(commentary)和追踪报道(follow-up).

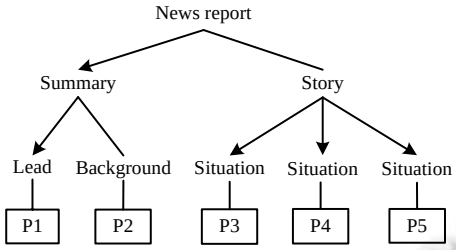

Fig.4 News schemata[34]

图4 新闻篇章的图式结构[34]

受这些语言学家研究成果的启发,本文综合语言学家的基本定义,为所有的叶子节点和非叶子节点定义了功能(function)属性,用来表示每个篇章单元在整个篇章中的作用和角色.如chtb_0094可用图5的功能语用结构树来表示.在这个功能语用结构树中,每个节点都表示其对应的节点的功能语用.例如,P1 简明扼要的描述了全文的主要内容,其功能就是全文的"导语";而 P1 和 P2 组合后的篇章单元是全文的总领,其功能就是全文的"总述".每个篇章的根节点表示全文的功能,那么对于新闻篇章来说,其根节点就是"新闻报道".

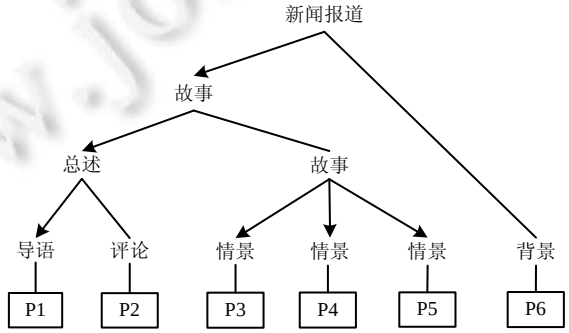

Fig.5 Macro discourse functional pragmatic structure of chtb_0094

图5 chtb_0094的宏观篇章功能语用结构

4.1 叶子节点

功能语用结构与逻辑语义结构的叶子节点相同,同样是自然分割的篇章段落.每个段落在篇章中所起的功能语用被标识为段落的一个属性,表示该段落在所在篇章关系中和整个篇章中的功能作用.

4.2 非叶子节点

在功能语用结构中,连接后的篇章单元的功能被标识为非叶子节点的一个属性,根据该节点在全文中的作用和篇章关系中的角色定义.每篇文章最后都形成一个完整的篇章结构树,本文将整个篇章的功能作为根节点的属性(新闻报道 news report).一个完整的篇章,其内容都是为全文的主旨服务的,形式衔接,内容连贯,每一个篇章单元的功能也将整个篇章紧密的连接在一起.每一个节点,无论是叶子节点还是非叶子节点,都在全文中发挥其功能作用,这些紧密连接的节点,才能组成一个完整的篇章.

4.3 功能节点定义

本文定义了18种功能类型,具体的功能节点定义如下.

- 新闻报道(news report):全文的根节点,所有的功能节点一层层向上连接,最终构成了一篇完整的新闻报道.
- 导语(lead):反映全文的中心话题,同时包含新闻的来源,即电头信息,如记者姓名、报道时间、报道地点等.一般是新闻报道的第 1 段.
- 总述(summary):全文的总述部分,一般出现在第 1 段,或者由文章的前几段组合而成.
- 次总述(sub-summary):文中局部部分的总述.例如,Sub-Summary 有时出现在 Story 的部分,表达局部总述的含义,相当于子话题.
- 总结(sumup):对前文或全文内容的重述或总结,一般出现在全文的最后一段,或部分段落之后.
- 情景(situation):支撑中心话题的细节描述,常常描述一个情节或事件.
- 故事(story):由 1 个或多个情景构成,表达完整的故事.
- 原因(cause):与结果(result)成对出现,指引起一定现象的现象.
- 结果(result):与原因(cause)成对出现,指由于原因的作用而引起的现象.
- 行为(behavior):与目的(purpose)成对出现,指为某一目的而执行的行为和动作.
- 目的(purpose):与行为(behavior)成对出现,指行为的所要达成的目的和效果.
- 陈述(statement):与举例(illustration)成对出现,用于描述事物、事件或观点.
- 举例(illustration):与陈述(statement)成对出现,通过列举事例来阐述陈述篇章单元所描述的事物、事件或观点.
- 背景(background):描述事件、事物的历史情况或现实环境等.
- 评论(comment):表达对前述事件、事物的判断、分析或看法.
- 补充(supplement):对前述事件、事物添加的说明、解释或额外内容.
- 对比(contrast):与前述单元对立的看法或事物,或与前述单元对立的另一个方面.
- 递进(progression):意义比前述单元更进一步,程度更深或意义更大.

如图 4 中 chtb_0094 的功能语用结构,其根节点即为该新闻报道(news report);P1 是全文的导语(lead);P2 评论了 P1,其功能是评论(comment);P3~P5 分别是对 P1 内容的细节描述,对应的功能是情景(situation);P3~P5 构成一个完整的故事(story);P1,P2 是全文的总述部分(summary);P6 的功能是全文的背景(background).

4.4 与RST和CDT的比较

本文定义的宏观篇章结构表示体系 MDT 与 RST 和 CDT 的最主要的区别是:本文关注的重点是宏观篇章结构,而 RST 和 CDT 的关注点主要是微观篇章结构.

(1) 基本篇章单元的定义:RST 中的 EDU 定义为子句或短语,CDT 中的 EDU 对应的是子句,以逗号作为标志;而 MDT 关注的是段落层以上的宏观篇章关系,因此,最基本的篇章单元对应的是作者在写作时按照其写作意图而自然分割的段落.

(2) 篇章关系定义:RST 对篇章关系的定义比较细,并且定义了每个关系 Nuclear 和 Satellite 对应的功能和要达到的目的,最初定义了 23 种篇章关系,RST-DT 标注时扩展成了 53 种单核关系和 25 种多核关系.CDT 中定义了 4 大类 17 种篇章关系.MDT 因分析对象是段落,一个段落表达的信息量比较大,太细致的分类无法适应,比如一个段落既阐述了事件发生的外部环境,也描述了事件的历史背景,此时如果使用 RST 的定义,就无法确定是 Circumstance 关系还是 Background 关系,因此,本文采用背景关系(background)来覆盖这两个定义.另外一些微观的篇章关系中宏观层面也不会出现,例如转折关系、条件关系、让步关系等,因此,MDT 定义了 3 大类 15 种篇章关系.RST、CDT 和 MDT 的篇章关系都是可扩展的,根据需要可以增加新的篇章关系定义.

(3) 主次关系定义:RST 中对应 Nuclear 和 Satellite,其判断标准与篇章关系直接相关.CDT 中对应主次关系,判断时需要考虑全局因素.MDT 的主次关系的判断依据主要是哪个篇章单元与篇章主题更相关,需要把握全局的内容和主旨.

(4) 篇章结构定义:根据定义,RST、CDT 和 MDT 都可以构建完整的篇章结构树.区别是 CDT 是微观篇章结构,MDT 是宏观篇章结构.RST 可以构建一棵完整的树.根据第 2 节宏观和微观统一的多层篇章结构的定义,MDT 和 CDT 结合,亦可以生成一棵完整的篇章结构树.

(5) 功能语用结构:RST 和 CDT 没有独立的功能语用结构定义,其内容涵盖在逻辑语义结构的定义中.MDT 中,我们根据 van Dijk 的宏观结构理论(macrostructure theory)中假拟新闻图式结构[34]和标注篇章的特点进行了定义.本文首次针对功能语用结构进行表示体系的定义,为后续篇章单元的功能语用研究奠定了坚实的基础.

为了更清晰地展示 3 种篇章结构体系的区别,本文将主要的区别罗列见表 3.

Table 3 MDT versus RST and CDT

表 3 MDT 与 RST、CDT 的对比

类别	RST	CDT	MDT
基本篇章单元	以子句或短语为基本篇章单元	以子句为基本篇章单元, 以逗号为分隔符标志	以自然分割的段落为 基本篇章单元
篇章关系	最初定义 24 种篇章关系,RST-DT 标注了 78 种篇章关系,可扩展	定义了 4 大类 17 种 篇章关系,可扩展	定义了 3 大类 15 种 篇章关系,可扩展
主次关系	对应“核心”和“卫星”,其判断 标准与篇章关系直接相关	判断时需要考虑当前关系 和上下文因素	判断依据主要是哪个篇章单元 与篇章主题更相关
篇章结构	可以构建完整的篇章结构树	可以构建完整的微观篇章结构树	可以构建完整的宏观篇章结构树
功能语用结构	没有独立的功能语用结构定义	没有独立的功能语用结构定义	有独立的功能语用结构定义

5 语料标注

大规模标注语料库的出现,带来了自然语言处理领域思维模式的转变,为基于统计的自然语言处理奠定了坚实的基础.根据第 1 节我们对相关工作的总结与分析,宏观篇章结构分析的语料资源目前还非常匮乏,尤其是在汉语篇章结构分析方面,目前尚未有宏观篇章结构语料.因此,构建宏观和微观统一的篇章结构语料库是非常有必要的,这不仅能够为篇章结构分析奠定良好的基础,还可以为自动文摘、机器翻译、信息抽取、问答系统等自然语言的应用提供有力的支持.

依据语料库构建的要求,可实现、可重复、高度一致,我们依据第 2 节~第 4 节关于宏观篇章结构表示体系的定义,进行了宏观汉语篇章结构语料的标注工作,并在标注的过程中,迭代地对宏观篇章结构表示体系的定义和标注准则进行修正,历时将近 1 年,完成了 720 篇新闻报道的标注工作,其来源是宾州篇章树库(Chinese treebank 8.0,简称 CTB 8.0).

篇章单元都不是孤立的,因此对于篇章的分析,无论是构建结构还是识别关系,都需要考虑整个篇章的信息.也就是说,不能单一地从篇章单元本身来判断结构和关系,而必须对篇章全文有一个整体的认识和理解,根据整个篇章的结构、内容和主旨来判断篇章结构和关系,有充分的理解之后才能进行有效的标注.

5.1 标注策略

本文标注小组在进行宏观篇章结构的标注时,首先需要通读篇章全文,并理解每一个段落描述的内容,全面理解篇章的主旨和主要内容.在标注中,采用自顶向下和自底向上相结合的标注策略.

(1) 自顶向下

针对每一篇新闻报道,首先根据全文的内容和主旨,找到最上层的关系,即判断第 1 层的切分位置,并逐层向下递归切分,直至每一个独立的篇章单元.例如,全文采用一种“先总述、再分述”的形式来描述,那么第 1 层的切分位置位于“总述”和“分述”之间,保证切分位置符合篇章的逻辑语义.

采用这种自顶向下的策略,是因为这种策略符合中文的阅读习惯,并且有利于在宏观上整体把握全局的内容和特征.

(2) 自底向上

同时,标注者需要根据段落的关联程度和相似程度来判断两个或多个段落是否需要优先组合.关联程度指的是相邻的段落是否论述同一个主体,上下段落之间是否存在较强的逻辑关系.相似程度指的是两个段落的长度、形式、内容是否有较强的相似度.如果两个段落之间存在较强的关联程度或相似程度,那么这些段落需要先行连接,并作为一个整体再与其他篇章单元连接.例如,连续的几个段落阐述一个事件的几个情节,长度、内容、形式都相似,而与其他段落有较大差异,那么这几个段落需要优先连接.

采用这种自底向上的局部策略可以帮助标注者将部分内容优先连接后当作一个整体来处理,避免出现逻辑上的错误拆分.如例 2 中的 P3~P5,从形式和内容上都较为相似,并且从全文语义上可以判断它们都是对 P1 内容的详细阐述,因此可优先使用“并列关系”进行连接.

5.2 标注方法

给定一篇生语料,标注者需要对语料进行文本阅读分析、篇章结构检测、篇章关系识别、主次关系判别、篇章结构树构建等操作,最终形成一棵完整的篇章结构树.完成逻辑语义视图的标注后,通过基于规则的逻辑语义到功能语用的转换,完成功能语用的自动标注.标注平台可将所有的标注结果最终保存为 XML 文件.为了保证标注一致性,还需要对标注结果进行数据验证和一致性计算.标注完成后,对标注结果进行数据统计和分析.具体处理流程如图 6 所示.

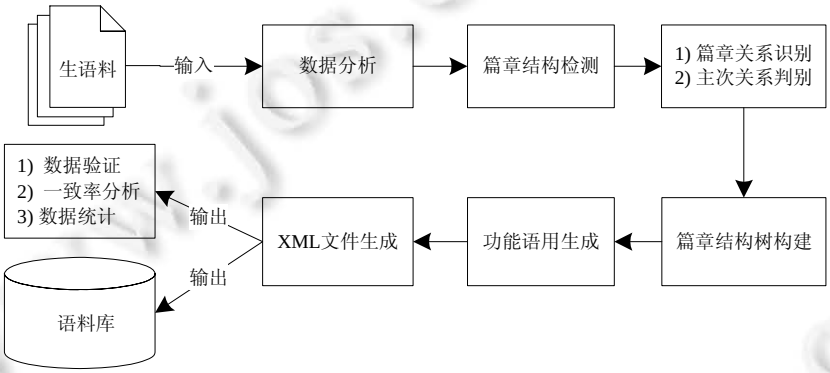

Fig.6 Processing flow of corpus annotating

图 6 语料标注处理流程

5.3 标注过程

我们将标注工作分为 3 个主要阶段.

- 第 1 阶段,尝试标注,历时 4 个月.从 CTB 8.0 中选取前 50 篇,3 名标注人员一同标注,经过逐篇的讨论,形成篇章结构表示体系的定义和初步的标注规范.去掉只有一个段落不能形成宏观篇章结构的 0045,第 1 阶段共完成 49 个篇章的标注.
- 第 2 阶段,优化规范,历时 3 个月.3 名标注人员独立标注,并进行小组讨论,同时讨论和修正表示体系定义和标注规范中不够明确的部分.第 2 阶段一共完成了 98 个篇章的标注.
- 第 3 阶段,批量标注,历时 3 个月.经过前两个阶段的标注工作,形成了较为规范的结构定义和较为稳定的标注质量.第 3 个阶段增加 3 名新的标注人员,6 名标注人员分为 3 个小组,每个小组新老标注人员搭配,独立标注,小组讨论.每次的标注工作分别计算一致率和 Kapp 值.每个标注周期,每个小组抽样 10%,由标注小组负责人进行交叉检查,并针对存有疑问的篇章进行组间讨论.该阶段共完成标注 573 篇.对每个小组的工作时间进行统计,计算平均标注效率,独立标注效率约 6.9 篇/时,小组讨论效率约 7.7 篇/时.

经过 3 个阶段,本文共完成 720 个篇章的标注工作.CTB 8.0 中的新闻语料有 753 个篇章,本文已完成其中所有较为规范的新闻文本的宏观篇章结构标注工作.

5.4 语料格式

本文将每一个篇章的标注结果保存在一个 XML 文件中.每一个 XML 文件包括 DISCOURSE、RELATION 和 TEXT 这 3 大部分,分别存储篇章信息、关系信息和段落信息.图 7 是例 2 所对应的 XML 文件.

```
〈?xml version="1.0",encoding="UTF-8"?〉
〈DOC〉
  〈DISCOURSE DiscourseTopic="海南洋浦开发区将动工兴建一批工业项目";
   Dateline="新华社海口一月六日电(记者柳昌林)"〉
    〈Lead〉海南洋浦经济开发区迎来工业建设高潮,5 个工业启动项目有的竣工投产,有的即将动工兴建.〈/Lead〉
    〈Abstract〉海南洋浦经济开发区迎来工业建设高潮,5 个工业启动项目:木浆厂、金岛精米加工厂、
          高速线材厂、橡木地板厂、浮法玻璃厂,有的竣工投产,有的即将动工兴建.〈/Abstract〉
  〈/DISCOURSE〉
  〈RELATION〉
    〈R ID="1",StructureType="逐层切分",Layer="1",RelationNumber="单个关系",RelationType="背景关系",
     ParagraphPosition="1...5|6...6",PrimarySecondary="1",ChildList="2",ParentId="−1",Function="News Report",
     RelationWeight="0"/〉
    〈R ID="2",StructureType="逐层切分",Layer="2",RelationNumber="单个关系",RelationType="解说关系",
     ParagraphPosition="1...2|3...5",PrimarySecondary="1",ChildList="3|4",ParentId="1",Function="Story",
     RelationWeight="0"/〉
    〈R ID="3",StructureType="逐层切分",Layer="3",RelationNumber="单个关系",RelationType="评价关系",
     ParagraphPosition="1...1|2...2",PrimarySecondary="1",ChildList="·",ParentId="2",Function="Summary",
     RelationWeight="0"/〉
    〈R ID="4",StructureType="并列切分",Layer="3",RelationNumber="单个关系",RelationType="并列关系",
     ParagraphPosition="3...3|4...4|5...5",PrimarySecondary="3",ChildList="·",ParentId="2",Function="Story",
     RelationWeight="0",/〉
  〈/RELATION〉
  〈TEXT〉
    〈P ID="1",ParagraphTopic="海南洋浦经济开发区迎来工业建设高潮.",ParagraphWeight="0",Function="Lead"〉.
     经过 5 年多的开发建设,海南洋浦经济开发区迎来工业建设高潮,5 个工业启动项目有的竣工投产,
     有的即将动工兴建.〈/P〉
    〈P ID="2",ParagraphTopic="开发区建设已由土地开发迈向工业项目建设的新阶段.",ParagraphWeight="0",
     Function="Comment"〉.洋浦经济开发区管理局局长王永春说,开发区招商工作取得突破性进展,
     开发区建设已由土地开发迈向工业项目建设的新阶段.〈/P〉
    〈P ID="3",ParagraphTopic="木浆厂已获国务院批准兴建.",ParagraphWeight="0",Function="Situation"〉.
     日前,洋浦开发区工业启动项目之一的 60 万吨木浆厂已获国务院批准兴建.这个全国最大规模的木浆厂
     将由新加坡亚洲浆纸业股份有限公司投资 12.83 亿美元兴建,年产漂白商品木浆 60 万吨.〈/P〉
    〈P ID="4",ParagraphTopic="金岛精米加工厂已竣工投产.",ParagraphWeight="0",Function="Situation"〉.
     首个工业启动项目——金岛精米加工厂已于去年底竣工投产.这个由澳门远东(泰国)集团公司与海南省
     粮油集团公司等联合投资三千万美元兴建的精米加工厂,采用国外 90 年代最先进的生产设备和工艺
     流程,年加工 30 万吨糙米,加工后的精米 70%出口.〈/P〉
    〈P ID="5",ParagraphTopic="其他 3 个启动项目也将动工投产.",ParagraphWeight="0",Function="Situation"〉.
     开发区的其他 3 个启动项目——高速线材厂、橡木地板厂、浮法玻璃厂也将动工投产.此外,开发区
     已确定和可能确定的工业项目还有 20 多个,包括油汽化工、钢铁厂、还原铁等,总投资约 70 亿美元.〈/P〉
    〈P ID="6",ParagraphTopic="洋浦情况介绍.",ParagraphWeight="0",Function="Background"〉.洋浦位于海南西部,是中国
     第一例由外商成片承包开发的工业开发区,享有目前国内最优惠、最开放的政策.过去 5 年间,土地开发商共投入
     40 亿港元用于电厂、区内主干道及地下管网、土地平整、邮电通讯等基础设施建设.(完).〈/P〉
  〈/TEXT〉
〈/DOC〉
```

Fig.7 Preservation form of corpus

图 7 语料的保存形式

DISCOURSE 部分存储篇章级的信息,是篇章全文的属性.标注内容包括:

(1) 篇章话题(discourse topic):存储全文的主题.

(2) 电头(dateline):存储电头信息(电头是电讯稿件发出单位、时间和地点的说明.电头多放在新闻稿件的开头,用括号或比较显著的字体区别于正文).

(3) 导语(lead):存储篇章的导语(导语一般是新闻的开头,以极其简洁的文字,写出消息中最重要的事实,提纲挈领,牵引全文,吸引读者),一般对应全文的第 1 段主要内容.

(4) 摘要(abstract):存储全文的摘要,由标注人员根据全文内容摘取.

RELATION 部分主要用于存储篇章结构和关系,是标注的主体.标注内容包括:

(1) 切分方式(structure type):存储篇章关系的切分方式,分为“逐层切分”(二元关系)和“并列切分”(对并列关系、二元或多元关系)两类.

(2) 切分层次(layer):自顶向下的切分层次,例如 Layer=“2”,表示该篇章关系的切分层次位于自顶向下的第 2 层.

(3) 篇章关系(relation type):存储对应的篇章关系,根据定义,包括 3 大类 15 种关系(篇章关系定义参见第 3.2 节).

(4) 关系个数(relation number):一个篇章关系为多个关系预留,分为“单个关系”和“多个关系”两类.

(5) 切分位置(paragraph position):存储篇章关系的切分位置,形如“2...3|4...4”,表示该关系为二元关系,前一个篇章单元由第 2 段和第 3 段组成,后一个篇章单元由第 4 段组成.

(6) 主次关系(primary secondary):存储该篇章关系的主次关系,根据定义,分别以“1”“2”“3”对应“主要-次要关系”“次要-主要关系”“同等重要关系”(主次关系定义参见第 3.3 节).

(7) 孩子节点(ChildList):存储下一个篇章关系的列表,最下层的孩子节点为空字符串.

(8) 父亲节点(parent Id):存储上一个篇章关系的编号,根节点的父亲节点为“−1”.

(9) 单元功能(function):存储篇章关系对应的非叶子节点的功能,如切分位置为“2...3|4...4”的篇章关系,其 Function=“Story”表示的是“2...4”的功能语用.

TEXT 部分用于存储篇章的段落信息(即叶子节点).标注内容包括:

(1) 段落话题(paragraph topic):存储该段落的主题信息.

(2) 段落功能(function):存储该段落对应的功能语用.

(3) 段落内容:存储整个段落对应的文本内容.

6 语料统计

6.1 标注一致性计算方法

为了验证标注质量,我们在第 2 和第 3 阶段分别进行一致性测试.一致性评估两名标注者的一致率.考虑到标注偶然一致的情况,本文还进行了 Kappa 值的计算.一致率 Agreement 的计算公式如公式(1)所示,Kappa 值的计算公式如公式(2)所示.公式(1)中,A 和 B 分别代表两名标注者.公式(2)中,$P(A)$表示标注一致的比例,$P(E)$表示偶然一致的比例.

$$Agreement=\frac{A\cap B}{A\cup B}\times 100\% \tag{1}$$

$$k=\frac{P(A)-P(E)}{1-P(E)} \tag{2}$$

本文采用的一致率计算方法参考 RST 的语料标注工作[49],并根据我们的标注内容做了适应性调整.具体的一致率计算方法,我们用例子加以说明.

首先,将树型的层次结构表示成边界标签的集合.如图 2 的篇章结构树,可将其结构表示成一组篇章单元边界的二元组[1,2],[3,5],[1,5],[1,6]构成.如另一名标注者的标注信息如图 8 所示,那么其标注的篇章结构可以用[1,2],[3,5],[3,6],[1,6]的二元组集合表示.

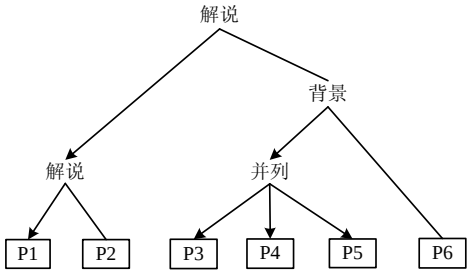

Fig.8 Another annotating discourse structure of chtb_0094

图 8 chtb_0094 的另一种篇章标注结构

如果出现多元关系,则需要将其转换为右链接的二元关系.比如,图 8 中 P3~P5 之间的并列关系,需要转换为 P4 和 P5 的并列关系;P3 与 P4,P5 的并列关系,如图 9 所示,其对应的边界二元组也从[3,5]转换为[4,5],[3,5].

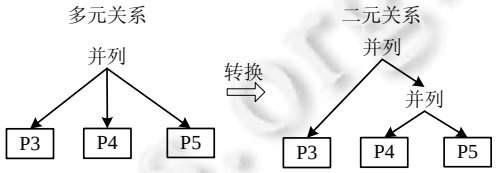

Fig.9 Convert multiple relationships to right-connected binary relationships

图 9 将多元关系转换为右连接的二元关系

其次,将两名标注者的标注信息汇总入一个标注集合表,作为待统计的元素,见表 4.

Table 4 Annotated sets of two annotators

表 4 两名标注者的标注集合

开始	结束	结构			主次			关系		
		标注者 *A*	标注者 *B*	是否一致	标注者 *A*	标注者 *B*	是否一致	标注者 *A*	标注者 *B*	是否一致
1	2	1	1	1	PS	PS	1	评价	解说	0
1	3	0	0	1	null	null	1	null	Null	1
1	4	0	0	1	null	null	1	null	Null	1
1	5	1	0	0	PS	null	0	解说	null	0
1	6	1	1	1	PS	PS	1	背景	解说	0
2	3	0	0	1	null	null	1	null	null	1
2	4	0	0	1	null	null	1	null	null	1
2	5	0	0	1	null	null	1	null	null	1
2	6	0	0	1	null	null	1	null	null	1
3	4	0	0	1	null	null	1	null	null	1
3	5	1	1	1	EI	EI	1	并列	并列	1
3	6	0	1	0	null	PS	0	null	背景	0
4	5	1	1	1	EI	EI	1	并列	并列	1
4	6	0	0	1	null	null	1	null	null	1
5	6	0	0	1	null	null	1	null	null	1
一致性		*Agreement*=13/15=86.67%			*Agreement*=13/15=86.67%			*Agreement*=11/15=73.33%		

在这个表格中,结构列的标注信息用“1”表示有结构标注,“0”表示没有结构标注.同时,没有结构标注的行所对应的主次和关系列值为“null”,表示此处不存在主次和关系的标注信息.利用该标注集合表的统计数据,使用公式(1)和公式(2)分别计算篇章结构、主次关系、篇章关系的一致率和 Kappa 值.如表 3 所示的标注结果,篇章结构的一致率为 86.67%,Kappa 值为 0.70;主次关系的一致率为 86.67%,Kappa 值为 0.70;篇章关系的一致率为 73.33%,Kappa 值为 0.492.

由此可以看出,主次关系和篇章关系一致性的计算需要基于正确的篇章结构,因此,其一致性低于篇章结构

的一致性;并且因为篇章关系的种类较多,如果标注者对于文章的理解有一定的偏差,基于不同的篇章结构标注出的篇章关系 Kappa 值将比较低.

6.2 语料标注质量

本文第 2 阶段标注一致性的计算结果见表 5(统计自标注者 *A* 和标注者 *B* 的标注数据).宏观的标注相对于微观的标注更为主观,因为标注过程中没有连接词这样的线索可以参考,并且段落间的关系相对于句内和句间的关系更为松散,因此标注的一致性不是很高,其中,篇章结构的标注一致率为 88.54%,Kappa 值为 0.771;主次关系的标注一致率为 80.67%,Kappa 值为 0.694;篇章关系的标注一致率为 83.05%,Kappa 值为 0.556.经过几轮迭代的标注和讨论,第 2 阶段的标注工作达到了较为稳定的一致性.

Table 5 Annotating consistency of Stage 2

表 5 第 2 阶段标注一致性

类别	一致率(%)	Kappa 值
篇章结构	88.54	0.771
主次关系	80.67	0.694
篇章关系	83.05	0.556

第 3 阶段有 6 名标注者参与标注工作,我们采用分组的形式来进行标注工作.为了避免因个人风格相似而造成的小组风格差异,我们在过程中也进行了多次小组轮换,并分别统计了各个工作周期两名标注者之间的标注一致性.该阶段的标注一致性计算结果见表 6.

Table 6 Annotating consistency of Stage 3

表 6 第 3 阶段标注一致性

标注者	平均长度	一致率(%)			Kappa 值		
		篇章结构	主次关系	篇章关系	篇章结构	主次关系	篇章关系
A+*D*	5.2	86.70	84.40	81.30	0.697	0.681	0.634
B+*E*	5.6	81.82	79.55	76.52	0.601	0.601	0.565
C+*F*	4.7	79.80	76.90	75.00	0.575	0.584	0.573
A+*E*	5.5	82.20	80.80	77.40	0.587	0.597	0.534
B+*C*	6	91.35	88.65	83.24	0.781	0.737	0.631
D+*F*	5	85.95	83.47	80.17	0.672	0.644	0.609
A+*F*	5.7	91.28	89.23	88.72	0.753	0.711	0.716
D+*E*	5.3	90.48	85.74	83.33	0.788	0.719	0.691
A+*B*	5.3	85.56	83.76	76.53	0.661	0.653	0.531
C+*D*	6.5	87.26	86.28	82.35	0.676	0.679	0.607
E+*F*	4.5	82.50	73.33	80.00	0.598	0.477	0.585
A+*D*	8.9	87.63	86.32	82.37	0.623	0.607	0.510
B+*E*	6	82.19	80.28	77.40	0.587	0.597	0.534
C+*F*	4.9	90.20	88.23	78.43	0.792	0.776	0.633
平均值	5.7	86.07	83.35	80.20	0.671	0.647	0.597

6.3 语料统计

本文已标注 720 篇文章,共有 3 985 个段落,形成 2 870 个篇章关系.平均每篇有 5.53 个段落,每篇最大段落数为 22,最小段落数为 2.总句子数为 8 319 句,平均段落长度为 2.1 个句子.共计 398 829 个字,平均每篇文档 554 个字.具体的统计数据见表 7.

本文对标注的篇章关系和篇章主次关系进行了数据统计.篇章关系方面,并列类关系共 1 253 个,占 43.66%;因果类关系共 414 个,占 14.43%;解说类关系共 1 203 个,占 41.92%.与并列类和解说类关系相比,因果类关系数据量偏少,数据集存在不平衡显现.主次关系方面,PS 关系共有 2 030 个,占 70.73%;SP 关系共有 80 个,占 2.79%;EI 关系共有 760 个,占 26.48%.与 PS 和 EI 关系相比,SP 关系的数据量很小,数据集存在较为严重不平衡的现象.统计数据见表 8.

Table 7 Basic statistic data of corpus

表 7 语料基础数据

统计项	统计值
篇章总数	720
段落总数	3 985
篇章关系总数	2 870
平均段落数(段落/篇章)	5.53
单个篇章最大段落数	22
单个篇章最小段落数	2
句子总数	8 319
平均句子数(句子/篇章)	11.55
总字数	398 829
平均字数(字数/篇章)	554

Table 8 Statistics of discourse relations and primary-secondary relations

表 8 篇章关系和主次关系统计

类别	篇章关系	PS	SP	EI	小计	百分比(%)
并列类	并列关系	0	0	634	634	22.09
	顺承关系	5	8	86	99	3.45
	递进关系	2	5	3	10	0.35
	对比关系	7	2	8	17	0.59
	补充关系	493	0	0	493	17.18
并列类小计		507	15	731	**1 253**	**43.66**
因果类	因果关系	16	24	9	49	1.71
	果因关系	93	3	7	103	3.59
	背景关系	237	0	0	237	8.26
	行为目的关系	12	0	2	14	0.49
	目的行为关系	7	2	2	11	0.38
因果类小计		365	29	20	**414**	**14.43**
解说类	解说关系	994	0	0	994	34.63
	总结关系	9	11	3	23	0.80
	评价关系	116	5	6	127	4.43
	陈述举例关系	38	1	0	39	1.36
	举例陈述关系	1	19	0	20	0.70
解说类小计		1158	36	9	**1 203**	**41.92**
主次关系	数量	**2 030**	**80**	**760**	2 870	
	百分比(%)	70.73	2.79	26.48	100.00	

每个类别的篇章关系的数量以及主次关系的分布情况见表 7.在 720 个篇章关系中,并列关系出现了 634 次,占 22.09%;解说关系出现了 994 次,占 34.63%.这两个关系一共占所有篇章关系的 56.72%,超过了整个语料的半数.由此现象可以推断,新闻报道类的文章常存在“总述-分述”这样的篇章模式,由 1 个或几个段落总领全文,再由一些段落分别从几个方面或几个角度进行详细阐述.这符合我们通常对于新闻类文章的认知,也符合新闻语料的写作框架.

统计数据同时显示了篇章关系和主次关系之间存在着很强的关系,例如,634 个并列关系均为 EI 关系,994 个解说关系全部是 PS 关系,补充、背景、陈述举例关系都是 PS 关系等.这种篇章关系和主次关系存在着极强关联性的现象,也是未来将要进一步分析和研究的方向之一.

针对功能语用结构的标注,统计数据见表 9.可以看出,大部分的新闻报道(news report)从基本形式上由总述(summary)和故事(story)组成,形成一种“总-分”的结构.新闻报道的内容上包含大量的事实描述,即故事(story)和情节(situation),而其他的信息,如评价(comment)、原因(cause)、结果(result)、行为(behavior)、目的(purpose)出现的较少.这些都符合新闻报道的特点.

Table 9 Statistics of functional pragmatic of discourse units

表 9 篇章单元功能语用统计

功能	数量	百分比(%)	功能	数量	百分比(%)
背景	237	3.46	目的	25	0.36
行为	25	0.36	结果	152	2.22
原因	152	2.22	情景	2 085	30.42
评论	127	1.85	陈述	59	0.86
对比	17	0.25	故事	1 441	21.02
举例	59	0.86	次总述	292	4.26
导语	638	9.31	总述	300	4.38
新闻报道	720	10.50	总结	23	0.34
递进	10	0.15	补充	493	7.19

6.4 实验验证

宏观篇章结构分析的任务包括篇章结构检测、主次关系判别、篇章关系识别和篇章结构树构建,其最终目标是构建一棵完整的宏观篇章结构树.其中,篇章结构检测是形成完整的篇章结构树的第 1 步,也是最重要的环节.因此,为了评估本文标注的语料的可计算性,本文给出了篇章结构检测的实验结果.

本文利用线性条件随机场作为基础分类器进行篇章结构检测和主次关系判别,为了减少相关任务之间的错误传递,构建了这两个任务的联合学习模型,并提出了一种基于整数线性规划的方法来实现全局优化.MCDTB 共有 8 863 个实例,其中正样例 3 261 个,负样例 5 602 个,本文采用 5 倍交叉验证来确保实验的客观性,使用准确率(accuracy)和宏观 $F1$ 值(macro-$F1$)来评价系统性能.

表 10 比较了运用不同的特征组合后,篇章结构检测和主次关系判别基础模型的性能.数据表明,组织结构特征对模型的贡献最大,结合了语义相似性特征后模型的性能得以提高,篇章结构检测和主次关系判别任务的性能分别达到 77.52%和 75.50%的准确率,相应的宏观 $F1$ 值分别达到 75.98%和 49.83%.在下面的联合模型实验中,用基础模型的最好性能进行比较.当应用了整数线性规划的约束条件时,篇章结构的推理准确率和宏观 $F1$ 分别达到 78.54%和 77.68%,比基础模型的最优性能分别提高了 1.02%和 1.70%;联合学习的方法同样提高了主次关系判别的性能,准确率和宏观 $F1$ 值分别提高了 0.51%和 1.86%.

Table 10 Exprimental results on structure identification and nuclearity recognition

表 10 篇章结构检测和主次关系判别实验结果

特征集合	篇章结构检测		主次关系判别	
	Accuracy	Macro-$F1$	Accuracy	Macro-$F1$
结构特征	76.13	74.46	74.46	49.17
语义相似度特征	74.57	73.32	66.68	37.70
结构特征+语义相似度特征	77.52	75.98	75.50	49.83
在最好性能上进行整数线性规划	78.54	77.68	76.01	51.69

7 总结与展望

本文针对宏观篇章结构分析目前研究较少的问题,分析了篇章结构分析的研究现状,提出了一种宏观和微观统一的篇章结构表示体系,并在宏观层面分别构建了篇章的逻辑语义结构和功能语用结构.基于本文构建的宏观篇章结构体系,标注了 720 篇规模的宏观篇章语料库.

理论体系的研究和建设是自然语言处理领域研究的基石,将直接影响到语料资源的质量和计算模型的性能.语料资源是基于统计的自然语言理解和分析的基础资源,构建大规模、高质量的语料资源具有较高的应用价值.因此,构建一个宏观和微观统一的理论体系,并建设相应的篇章结构语料资源,为后续的研究工作打好了坚实的研究基础.在此基础上,才可以进一步深入研究篇章结构分析的计算模型建设,从而提高篇章分析的整体性能,进而为自然语言相关应用提供服务,以期得到更高的系统性能,推动人工智能领域应用的飞速发展.

本文现有标注的语料规模还不够大,后续我们将进一步完善标注规则,扩大语料规模,并考虑其他领域的宏

观篇章结构特点进行标注.对数据不平衡问题,进一步分析其出现的原因及应对策略.在此基础上,将进行篇章结构分析的计算方法和计算模型的研究,并最终构建一个 End-to-End 的宏观篇章结构分析平台.

References:

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

[2] Ferreira R, de Souza Cabral L, Freitas F, Lins, RD, de França Silva G, Simske SJ, Favaro L. A multi-document summarization system based on statistics and linguistic treatment. Expert Systems with Applications, 2014,41(13):5780−5787.

[3] Cohan A, Goharian N. Scientific article summarization using citation-context and article's discourse structure. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015. 390−400.

[4] Meyer T, Popescu-Belis A. Using sense-labeled discourse connectives for statistical machine translation. In: Proc. of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation. Stroudsburg: ACL, 2012. 129−138.

[5] Guzmán F, Joty S, Màrquez L, Nakov P. Using discourse structure improves machine translation evaluation. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014. 687−698.

[6] Peldszus A, Stede M. Joint prediction in MST-style discourse parsing for argumentation mining. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2015. 938−948.

[7] Presutti V, Draicchio F, Gangemi A. Knowledge extraction based on discourse representation theory and linguistic frames. In: Proc. of the Int'l Conf. on Knowledge Engineering and Knowledge Management. Berlin: Springer-Verlag, 2012. 114−129.

[8] Zou B, Zhou G, Zhu Q. Negation focus identification with contextual discourse information. In: Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014. 522−530.

[9] Liakata M, Dobnik S, Saha S, Batchelor C, Rebholz-Schuhmann D. A discourse-driven content model for summarizing scientific articles evaluated in a complex question answering task. In: Proc. of the Conf. on the Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2013. 747−757.

[10] Chu X, Zhu Q, Zhou G. Discourse primary-secondary relationships in natural language processing. Chinese Journal of Computers, 2017,40(4):842−860 (in Chinese with English abstract).

[11] Li Y, Feng W, Sun J, Kong F, Zhou G. Building Chinese discourse corpus with connective-driven dependency tree structure. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014. 2105−2114.

[12] Xue N, Chiou FD, Palmer M. Building a large-scale annotated Chinese corpus. In: Proc. of the 19th Int'l Conf. on Computational Linguistics. Vol.1. Stroudsburg: ACL, 2002. 1−8.

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

[14] Halliday MAK, Hasan R. Cohesion in English. Longman, 1976.

[15] Hobbs JR. On the coherence and structure of discourse. In: Proc. of the Center for the Study of Language and Information. 1985. 1−36.

[16] Hobbs JR. Coherence and coreference. Cognitive Science, 1979,3(1):67−90.

[17] Mann WC, Thompson SA. Relational propositions in discourse. Discourse Processing, 1986,9(1):57−90.

[18] Mann WC, Thompson SA. Rhetorical structure theory: A theory of text organization. Technical Report, ISI/RS-87-190, Information Sciences Institute, University of Southern California, 1987.

[19] Mann WC, Matthiessen C, Thompson SA. Rhetorical structure theory and text analysis. In: Proc. of the Discourse Description: Diverse Linguistic Analysis of a Fund-raising Text. 1992. 39−78.

[20] Xue N. Annotating discourse connectives in the Chinese treebank. In: Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2005. 84−91.

[21] Prasad R, Dinesh N, Lee A, Miltsakaki E, Robaldo L, Joshi AK, Webber BL. The Penn discourse treebank 2.0. In: Proc. of the Language Resources and Evaluation Conf. Berlin: Springer-Verlag, 2008. 2961−2968.

[22] Zhou Y, Xue N. PDTB-style discourse annotation of Chinese text. In: Proc. of the Association for Computational Linguistics. Stroudsburg: ACL, 2012. 69−77.

[23] Grosz BJ, Sidner CL. Attention, intentions, and the structure of discourse. Computational Linguistics, 1986,12(3):175−204.

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

[25] Wu WZ, Tian XL. The Chinese Sentence Group. Beijing: The Commercial Press, 2000 (in Chinese).

[26] Xing FY. Research on Chinese Complex Sentence. Beijing: The Commercial Press, 2001 (in Chinese).

[27] Yao SY. A research on the collocation of the relation markers of Chinese compound sentences and some relevant explanation [Ph.D. Thesis]. Wuhan: Central China Normal University, 2006 (in Chinese with English abstract).

[28] Li YC. Research of Chinese discourse structure representation and resource construction [Ph.D. Thesis]. Suzhou: Soochow University, 2015 (in Chinese with English abstract).

[29] Hoey M. On the Surface of Discourse. Buckley: George Aller, and Unwin Publisher, Ltd., 1983. 93−129.

[30] Martin JR, Rose D. Working with Discourse: Meaning Beyond the Clause. London: Continuum, 2003.

[31] van Dijk TA. Macrostructure: An Interdisciplinary Study of Global Structures in Discourse, Interaction, and Cognition. Hillsdale: Lawrence Erlbaum Associates, Inc., 1980.

[32] van Dijk TA. Text and Context: Explorations in the Semantics and Pragmatics of Discourse. London: Longman, 1977.

[33] van Dijk TA, Kintsch W. Strategies of Discourse Comprehension. New York: Academic Press, 1983.

[34] van Dijk TA. Handbook of discourse analysis. In: Proc. of the Discourse and dialogue. Academic Press, 1985.

[35] van Dijk TA. News as Discourse. Hillsdale: Lawrence Erlbaum Associates, Inc., Publishers, 1988.

[36] Carlson L, Marcu D, Okurowski ME. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: Proc. of the Current and New Directions in Discourse and Dialogue. Springer Netherlands, 2003. 85−112.

[37] Prasad R, Miltsakaki E, Dinesh N, Lee A, Joshi A, Robaldo L, Webber BL. The Penn discourse treebank 2.0 annotation manual. 2007. https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf

[38] Yue M. Rhetorical structure annotation of Chinese news commentaries. Journal of Chinese Information Processing, 2008,22(4): 19−23 (in Chinese with English abstract).

[39] Soricut R, Marcu D. Sentence level discourse parsing using syntactic and lexical information. In: Proc. of the 2003 Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT). 2003. 149−156.

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

[41] Joty S, Carenini G, Ng RT, Mehdad Y. Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In: Proc. of the 51st Annual Meeting of Association for Computational Linguistics. Stroudsburg: ACL, 2013. 486−496.

[42] Joty S, Carenini G, Ng RT. A novel discriminative framework for sentence-level discourse analysis. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012. 904−915.

[43] Feng VW, Hirst G. A linear-time bottom-up discourse parser with constraints and post-editing. In: Proc. of the 50th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2014. 511−521.

[44] Ji Y, Eisenstein J. Representation learning for text-level discourse parsing. In: Proc. of the 52th Annual Meeting of the Association for Computational Linguistics (ACL). 2014. 13−24.

[45] Sporleder C, Lascarides A. Combining hierarchical clustering and machine learning to predict high-level discourse structure. In: Proc. of the 20th Int'l Conf. on Computational Linguistics. 2004. 43−49.

[46] Chu X, Wang Z, Zhu Q, Zhou G. Recognizing nuclearity between Chinese discourse units. In: Proc. of the 19th Int'l Conf. on Asian Language Processing. IEEE, 2015. 197−200.

[47] Fairclough N. Media Discourse. London: Edward Arnold, 1995.

[48] Bell A, Garrett PD. Approaches to Media Discourse. Oxford: Wiley-Blackwell, 1998.

[49] Marcu D, Amorrortu E, Romera M. Experiments in constructing a corpus of discourse trees. In: Proc. of the ACL'99 Workshop on Standards and Tools for Discourse Tagging. 1999. 48−57.

附中文参考文献:

[10] 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.

[25] 吴为章,田小琳.汉语句群.北京:商务印书馆,2000.

[26] 邢福义.汉语复句研究.北京:商务印书馆,2001.

[27] 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].武汉:华中师范大学,2006.

[28] 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,2013.

[38] 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.

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

徐昇(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.

奚雪峰(1978－),男,博士,副教授,CCF 专业会员,主要研究领域为自然语言理解,人机智能交互,机器学习.

朱巧明(1963－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为自然语言理解,信息抽取,信息检索,知识图谱.

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.

周国栋(1967－),男,博士,教授,博士生导师,CCF 杰出会员,主要研究领域为自然语言理解,机器翻译,信息抽取,信息检索,机器学习.