计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2017.20170348
54(8):1833-1852,2017

汉语篇章微观话题结构建模与语料库构建

奚雪峰[1,2,3] 褚晓敏[1] 孙庆英[1] 周国栋[1]
[1](苏州大学计算机科学与技术学院 江苏苏州 215000)
[2](苏州科技大学计算机科学与工程系 江苏苏州 215009)
[3](苏州市虚拟现实智能交互及应用技术重点实验室 江苏苏州 215009)
(xfxi@mail.usts.edu.cn)

Corpus Construction for Chinese Discourse Topic via Micro-Topic Scheme

Xi Xuefeng[1,2,3], Chu Xiaomin[1], Sun Qingying[1], and Zhou Guodong[1]
[1](*School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu* 215000)
[2](*Department of Computer Science and Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu* 215009)
[3](*Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou, Suzhou, Jiangsu* 215009)

Abstract Currently discourse topic structure analysis is the fundamental research of natural language understanding. Due to the lack of a large number of high-quality discourse corpus resources, which are suitable for Chinese discourse analysis, it has seriously restricted the research of the relevant discourse topic computing models. In order to solve the above problems, we firstly study the theoretical representation system of Chinese discourse topic structure. From the theme-rheme theory, theory of English rhetorical structure and Pennsylvania discourse treebank system, research of Chinese complex sentence and sentence group, combined with Chinese characteristics, we propose a Chinese discourse micro-topic scheme based on theme-rheme theory and construct a Chinese discourse topic structure representation model based on the topic chain. Then, on the basis of the above, we adopt the top-down and backward search annotation strategy and the combination of the human machine and the corpus annotation method to construct the Chinese discourse topic corpus (CDTC). Moreover, we carry out a detailed statistical analysis of the CDTC which contains a total of 500 documents. Compared with the OntoNotes corpus and the generalized topic structure theory, this micro-topic scheme representation model has some advantages in theory and is consistent with the Chinese characteristics. Finally, the consistency test shows that CDTC can fully reflect the difficulty of Chinese discourse topic analysis, and can provide support for the relevant research.

Key words discourse topic structure; theme-rheme theory; thematic progression; topic chain; corpus construction

摘 要 篇章话题结构分析是自然语言理解的前沿基础,而大规模高质量的适用于汉语篇章分析的语料资源缺乏,严重制约了相关篇章话题计算模型的研究.针对上述问题,首先研究了汉语篇章话题结构的

收稿日期:2017-05-23;**修回日期**:2017-06-21
基金项目:国家自然科学基金项目(61331011,61673290,61472264)
This work was supported by the National Natural Science Foundation of China (61331011, 61673290, 61472264).
通信作者:周国栋(gdzhou@suda.edu.cn)

理论表示体系.分析了主述位理论、英语修辞结构理论和宾州篇章树库体系的优势,结合汉语复句句群理论以及汉语自身特点,提出了一种基于主述位理论的汉语篇章微观话题结构表示方式,并借助微观话题链构建了汉语篇章话题结构表示体系.随后,在此基础上,采用自顶向下、后向搜索的标注策略和人机结合的语料库标注方式,构建了基于篇章微观话题表示体系的汉语篇章话题结构语料库(Chinese discourse topic corpus, CDTC).CDTC 共包含 500 个文档,对其进行了详细统计分析并展示了语料库的标注情况.与宾州篇章树库体系、广义话题结构理论的对比表明,所提篇章微观话题结构表示体系在理论上具有一定的优越性,并且符合汉语特点;一致性检验表明 CDTC 能够充分体现汉语篇章话题分析问题本身的难度,并能够为相关研究提供语料资源支持.

关键词 篇章话题结构;主位-述位理论;主位推进;话题链;语料库构建

中图法分类号 TP391

让机器准确地理解自然语言文本篇章主题,甚至能够理解篇章作者想要表达的意图,是人工智能发展的重大挑战任务之一.而在当前自然语言理解的研究工作中,篇章话题结构分析是前沿核心基础,主要任务是从篇章整体层次上分析篇章话题结构及其组成单元之间的语义关系,并利用上下文理解篇章.

然而由于大规模高质量的适用于汉语篇章分析的标注语料严重缺乏,制约了相关篇章话题计算模型的研究,近年来汉语篇章语料库资源建设已经逐渐成为研究重点.尽管已有研究者或基于英语篇章分析理论体系,或基于汉语的复句、句群理论和广义话题结构理论,对汉语篇章话题结构分析资源库展开了有益的探索[1-4],然而总体来看,汉语篇章话题结构语料库构建研究依然较为匮乏,这使得面向汉语的篇章话题结构研究受到极大的制约.

基于此,我们面向汉语篇章话题结构开展了针对性研究.基于主述位理论,提出了一种篇章微观话题结构形式化表示模型,并基于该模型完成了篇章话题结构语料库构建.

1 相关研究

1.1 汉语篇章话题结构分析

在汉语篇章话题研究方面,赵元任[5]首先在汉语结构分析研究中引入话题(topic)概念,他采用了"话题"和"说明"来阐释汉语的"主语"和"谓语"结构.曹逢甫[6]则强调了话题的篇章本性.在汉语篇章中,通过采用话题的代词化和省略形式,可以把话题的语义范围拓展到小句之外,而这种方式恰好有助于构建话题的链结构,体现篇章的衔接性;进一步,曹逢甫提出了汉语话题链(topic chain)的概念,并研究了在控制小句连接过程中,话题链所起到的作用.话题链的形成主要依赖各种指代回指(anaphor)形式,即零形回指(zero anaphor, ZA)、代词回指(pronoun anaphora, PA)和名词回指(nominal anaphor, NA)的选择方法.屈承熹[7]综合分析了已有研究成果,将话题链定义为"一组以零回指 ZA 形式的话题连接起来的小句",从而提供了较强的可计算性.

话题链在篇章结构分析中的独特作用,不仅在汉语篇章分析中存在,而且在英语篇章中也有类似效果.基于小规模人工标注的汉英篇章并行语料库,刘礼进对比研究了在宏观语义结构描述上,汉英篇章中存在的话题链所表现出的功能差异性[8].

王建国也分析了汉英篇章中话题链的不同特点,并拓展了话题链的定义,将其描述为"由同一话题引导的系列语句",这把话题链的作用范围延伸到句群和篇章层面[9].周强和周骁聪[10]定义了话题评述关系集合,结合关联词语及已有的连贯形式描述机制,构建了一种新的话题链描述形式.

汉语与英语相比有很大不同,就篇章结构而言,从基本篇章单元、篇章结构的组织、篇章关系的分类,到连接词的形式与分布均有所不同,因此面向英语篇章结构分析的修辞结构理论和宾州篇章树库体系并不能直接套用到汉语篇章结构分析应用中.

就汉语而言,具有"本土特征"的复句句群理论虽然着眼点不是篇章理论,但是徐赳赳[11]对比研究了复句句群理论和 RST 理论,发现两者研究的对象、内容、方法及其表现形式等都有相通之处,因而推断在汉语篇章分析层面,复句句群理论应该还有很大的潜在应用价值.

此外,宋柔等人针对汉语篇章话题结构进行了比较深入的研究,提出了广义话题结构(generalized

topic structure, GTS)的概念和相应的表示方法[1-2,12].依据这一理论,他们以标点句为基本篇章单位,开展了汉语篇章的话题结构标注工作.这一研究成果是汉语篇章分析领域的一项开创性工作.

相对于西方语言(特别是英语)篇章分析的长期研究,汉语篇章话题分析的研究刚刚起步,目前主要处于理论体系探索和语料库资源建设阶段.

1.2 汉语篇章结构语料库及计算模型

由于大规模高质量的适于汉语篇章分析的标注语料严重缺乏,制约了相关篇章话题计算模型的研究,近年来建立汉语篇章语料库资源日益成为研究者关注的焦点.相关工作主要包括2类:1)在参考RST和PDTB体系的基础上,结合汉语复句和句群理论的研究成果,对汉语篇章结构的标注体系进行探索;2)针对汉语篇章话题结构,开展相应语料库建设实践.

第1类代表性工作有:乐明[13]根据RST理论,面向汉语篇章结构,结合汉语句群和复句理论开展了标注探索,主要工作包括以标点符号为边界,定义了篇章修辞结构分析的基本单元;定义了47种汉语修辞关系集用于区分核心单元;定义了篇章结构标注的具体规则;在此基础上,选取来自大陆主要媒体中的97篇财经评论文章开展了修辞结构标注,探索了中文篇章分析中采用RST的可行性.此外,Xue[14]分析了汉语中篇章连接词的分布情况,并对汉语篇章连接词的意义消歧和变形等问题进行了探讨,他采用类PDTB标注体系,面向中文树库篇章连接词标注问题,尤其是显式连接词标注开展了标注实践及相关研究.在此基础上,Zhou等人[15]在PDTB标注体系下对来自中文树库的98个文件进行了标注,并对汉语和英语在该体系下的差异进行了分析.汉语句子中由于缺省较多连接词,因此无法直接采用面向英语的PDTB体系开展相关研究.Huang等人[16]提出一种弹性的汉语篇章结构标注框架并完成了网上标注系统的开发;结合此标注框架及PDTB体系,标注者完成了隐式或显式、跨句及句内等篇章关系,以及情感信息的标注.张牧宇等人选取LDC发布的OntoNotes 4.0中的1096篇汉语文本按照PDTB体系进行了分句、复句和句群3个层次的篇章关系的标注[17].标注内容包括显式篇章关系的关系连接词、关系元素和关系类别信息;以及隐式关系的可插入的连接词和篇章关系类别信息.他们将篇章关系分为时序、因果、条件、比较、扩展和并列6类,标注的关系连接词共有1273个.

作者所在的苏州大学自然语言课题组结合PDTB和RST体系的优势,在充分考虑汉语篇章特点的基础上,将基本篇章单位和连接词分别采用树结构的叶子节点及中间节点的形式加以表示.高级篇章单位相对分层:最底层由各基本篇章单元组成;组合底层的篇章单位,构建次高级的篇章;重复组合,不断产生更高级的篇章单位,最终将汉语篇章修辞结构表示成一棵篇章结构树[18].在此方案指导下,该课题组标注完成了中文树库上500篇文章的篇章修辞结构,其中涉及基本篇章单位、篇章结构边界、篇章连接词、篇章分层关系及主次篇章单位等.

第2类代表性工作是宋柔课题组基于他们提出的广义话题结构的概念,把标点句看作基本篇章单位,开展了汉语篇章的话题结构标注工作,已标注了《围城》、《鹿鼎记》和其他语料(涉及章回小说、现代小说、百科全书、法律法规、散文、操作说明书等不同语体),共约40万字,其数据仍在修订整理中.其中,《鹿鼎记》第1回的广义话题结构标注及其说明已经在网上公开发布①.

2 基于主述位理论的汉语篇章微观话题结构

形式化表示体系是语料库资源建设的基础.结合主述位理论[19]、汉语复句理论[20]、广义话题结构理论[21]等的研究,我们采用一种主述位形式表示汉语的篇章话题结构,并基于主位推进模式构建汉语的篇章话题联接关系体系,用以指导构建一个结构表示清晰、便于扩展对比、标注统一的汉语篇章话题结构语料库.

为便于说明问题,我们给出例1及其篇章话题结构的可视化表示如图1所示.

例1. a[李四]T_1 比较年轻,‖b[]T_2〈而且〉工作经验也不足,‖‖c[]T_3 学历也又不高,|d 但是[]T_4 不论做啥事情,‖‖e[他]T_5 都认真负责,‖f[所以,领导]T_6 非常器重他.

从图1中(b)部分内容可见,例1由多个篇章基本单元组成,各单元之间通过语义衔接关系相连接,为进一步开展篇章话题分析提供了表示基础.

面向图1中所示的篇章话题结构表示形式,我们从形式化表示及可计算角度给出定义.

① http://clip.blcu.edu.cn/

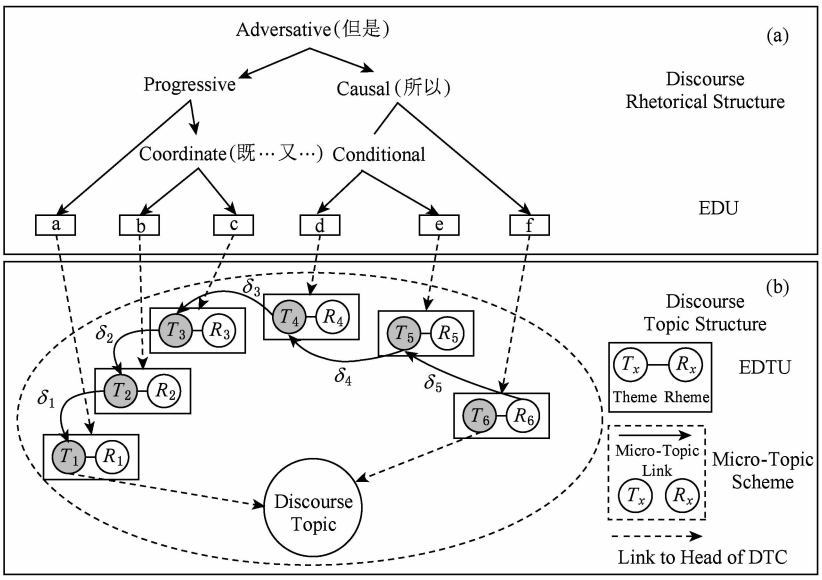


Fig. 1 Chinese discourse topic structure visualization representation for example 1

图 1 例 1 的汉语篇章话题结构可视化表示体系

定义 1. 篇章基本话题(elemental discourse topic unit, EDTU). 是最小的独立表达意图性的单位,通常是一个含主谓的独立句子.

例 2. 1) 两名俄罗斯航天员进入航天飞机.

2) 两名俄罗斯航天员进入航天飞机开始准备升空.

例 2 中 1)表示一个基本话题,主语是"两名俄罗斯航天员",谓语是"进入(航天飞机)";而例 2 中 2)则包含了 2 个基本话题,主语相同,谓语分别是"进入(航天飞机)"和"开始(准备升空)".

在图 1 表示例 1 的汉语篇章话题结构中,篇章基本话题共有 6 个,分别表示例 1 中的 a～f 标注段. 这里所提到的篇章基本话题结构(EDTU),从形式上与有关篇章修辞结构中所标注的篇章子句(EDU,图 1 中(a)部分)是一致的[18],这也有利于开展篇章修辞结构与篇章话题结构的统一联合研究.

篇章微观话题结构以篇章基本话题(EDTU)为组成部分,借鉴主述位理论,给出篇章微观话题结构定义.

定义 2. 篇章微观话题结构(micro-topic scheme, MTS). 是一个四元组,

$$MTS=(S_n,S_{n+1},V,\delta_n),$$

其中,$S_n\in T\cup R$,$S_{n+1}\in T\cup R$,T 为一个篇章中的篇章基本话题(EDTU)的主位(theme)集合;V 为连接成立的置信函数,$V(S_n,S_{n+1})\in[0,1]$;R 为同一个篇章中的篇章基本话题(EDTU)的述位(rheme)集合;$\delta_n\in\Gamma$,Γ 为同一个篇章中的微观话题联接(micro-topic link)的集合.

定义 3. 置信函数 V. $\forall V$,当且仅当 $V\geqslant Threshold$,$0\leqslant Threshold\leqslant 1$,$\delta_n$ 才能成立.

这里 $Threshold$ 表示连接成立的阈值,δ_n 成立即表示 MTS 成立.

例 3. 1) 这份文件是玛利亚留下的;

2) 她刚刚走开.

例 3 中,篇章基本话题 1)与篇章基本话题 2)之间即通过微观话题联接构成一个篇章微观话题结构. 其中,"是玛利亚留下的"是基本话题 1)中的述位,而"她"则是基本话题 2)中的主位.

有关主位、述位以及微观话题联接的定义见定义 4～定义 6.

定义 4. 篇章微观话题结构中的主位(theme). 是指包含在一个篇章基本话题(EDTU)之中的谓词前面的成分,一般包含主语;篇章基本话题(EDTU)中剩余部分,即为述位(rheme).

汉语重意合,在子句中会大量出现缺省主语(或宾语等)的情况,因此也带来了包含主语的主位(或包含宾语的述位)的缺省.

定义 5. 隐式主位. 一个篇章基本话题中缺省的主位,称为隐式主位;缺省的述位称为隐式述位.

需要特别说明的是,由于自然语言中句子包含陈述句、一般疑问句、复杂疑问句、反问句、祈使句等多种句型,相应的主述位定义也有所不同. 为降低初

始研究的复杂性,定义 4～5 中的主述位概念,均限定在陈述句范畴内.后续将对其他句型分类展开形式化定义研究.

例 3 中,基本话题 2)的主位"她"与基本话题 1)的述位中的"玛利亚"形成指代关系.这里的指代关系即为一种语义关联,形成微观话题联接(micro-topic link).

定义 6. 微观话题联接(micro-topic link,MTLink).是一种上下文篇章基本话题(EDTU)内主述位之间语义关联的可信度表示,体现篇章之间的衔接特性,主要包含照应(指代)、省略、替代、重复、同义、反义、具体抽象化(下义转上义)、抽象具体化(上义转下义)、局部整体化、整体局部化、搭配共 11 种类型,形式化为

$$MTL=(CT,CV),$$

其中,CT 是一种衔接类型,$CT\in${照应、省略、替代、重复、同义、反义、具体抽象化、抽象具体化、局部整体化、整体局部化、搭配};CV 是衔接类型成立的可信度,取值为实数,其值区间表示为 $0\leqslant CV\leqslant 1$.

下面分别对 11 种衔接类型加以说明:

1) 照应.指的是一个主述位作为另一个篇章基本话题(EDTU)中主述位的参照点,如例 4 中的人称代词"他"指前面出现的"彼得".

例 4. 彼得有一个妻子,非常爱他.

2) 省略.指的是把一个篇章基本话题(EDTU)中的主述位忽略不计,从而避免重复.这种衔接类型有利于突出新信息,形成上下紧凑的语篇关系.如例 5 中,"看到一只猫前"省略了"我".

例 5. 我早上出门,看到一只猫.

3) 替代.指的是用替代词去取代篇章基本话题(EDTU)中的主述位,替代词的语义来自于所替代的成分.

4) 重复.指的是篇章基本话题(EDTU)中的主述位多次出现,如例 6 中的"熊".

例 6. 安哥拉碰到了一只熊,这只熊显然非常饥饿.

5) 同义.指的是关联上下 2 个篇章基本话题(EDTU)中的主述位是一对同义词.

6) 反义.指的是关联上下 2 个篇章基本话题(EDTU)中的主述位是一对反义词.

7) 具体抽象化.指的是存在关联关系的两个篇章基本话题(EDTU)中的主述位,上一个主述位是下一个主述位的具体表示(或者是子类),下一个主述位是上一个主述位的抽象表示(或者是父类).如例 7 中,上一个主述位所标注的"哺乳动物"属于"动物"的一种,"动物"是"哺乳动物"的父类.

例 7. 哺乳动物/是动物的一种,这里的动物/大多属于能够自主移动的生命体.

8) 抽象具体化.指的是存在关联关系的 2 个篇章基本话题(EDTU)中的主述位,上一个主述位是下一个主述位的抽象表示(或者是父类),下一个主述位是上一个主述位的具体实例表示(或者说是子类).如例 8 中,上一个主述位所标注的"哺乳动物"是"马"的抽象表示,"马"是一种具体的"哺乳动物".

例 8. 哺乳动物/存在一类食草动物,马/就是这类动物的典型代表.

9) 整体局部化.指的是下一个篇章基本话题中的主述位是上一个篇章基本话题主述位的局部表示.如下例 9 中的"脸","身"和"手",可以同上文提到的"一个老头"形成局部与整体的语义关系.

例 9. 前面走来一个老头,满脸皱纹,身披破棉袄,手中拿着个搪瓷碗.

10) 局部整体化.指的是下一个篇章基本话题中的主述位是上一个篇章基本话题主述位的整体表示.如例 10 中的"这辆自行车"可以同上文中提到的"轮毂"、"车身"形成整体与局部的语义关系.

例 10. 轮毂/变形了,车身架子/断裂了,这辆自行车/基本报废了.

11) 搭配.指的是词汇同现,即一组语义上有联系的词汇关联上下篇章基本话题(EDTU)结构中的主述位.例如 2 组词:(冰天雪地,白色)和(夜晚,星星).

在图 1 表示的例 1 的汉语篇章话题结构中,篇章微观话题结构(MTS)共有 5 个,分别以微观话题联接(图 1 中(b)部分的箭头)相关联,可以表示为 (T_1,T_2,V,δ_1),(T_2,T_3,V,δ_2),(T_3,T_4,V,δ_3),(T_4,T_5,V,δ_4),(T_5,R_6,V,δ_5),其中置信函数 V 的取值都是大于阈值的.

定义 7. 篇章话题结构(discourse topic structure,DTS).由 $n(n\geqslant 1)$ 个篇章微观话题结构(MTS)组成,且篇章微观话题结构(MTS)之间也通过篇章微观话题联接(MTLink)相关连.

实质上,篇章话题结构是一种递归定义,可以表示如下规则为:

1) 篇章微观话题结构是篇章话题结构;

2) 通过篇章微观话题联接(MTLink)相关联的两个篇章话题结构也是篇章话题结构;

3) 篇章话题结构,当且仅当有限次使用规则 1 和规则 2 所构成.

定义 8. 篇章话题链. 在一个篇章话题结构中，多个相关联的篇章微观话题联接(MTLink)构成了一个篇章话题链(discourse topic chain, DTC).

在图 1 表示的例 1 汉语篇章话题结构中，δ_1，δ_2，δ_3，δ_4，δ_5 构成了一个篇章话题链.

主位推进理论中的主位推进模式直观地反映了篇章话题演变关系，将其应用于汉语篇章话题链的识别即可构建一个完整的篇章话题结构体系. 本文把篇章话题演变关系作为一个独立模块进行研究，资源标注部分同时也涉及话题结构体系. 两者的关系在于，在标注资源上对篇章基本话题进行话题链的识别，即可获得篇章话题动态演变模型. 这里需要研究的问题是如何对话题链进行形式化表示，从而获得一个逻辑性强，并且应用广泛的篇章话题结构体系. 鉴于此，我们提出一个基于微观话题结构的主位推进模式(MTS-TPs).

定义 9. 微观话题主位推进模式. 在一个篇章微观话题结构中，根据微观话题联接(MTLink)所连接的上下篇章基本话题中的不同主位或述位，可以构成不同的微观话题主位推进模式(MTS-based thematic progression patterns, MTS-TPs)，其实质也是一种微观话题结构. 形式化表示为 $MTS-TP_s \in \{MTS\}$.

与传统主位推进模式表示不同的是，我们在判断上下子句之间的主述位关系时，不仅包含传统主述位语义相等关系，而且还提出了包含其他具有篇章之间衔接关系的微观话题联接(MTLink)的概念(见定义 6)，即主述位之间只要形成微观话题联接，上下句之间的关联关系就能成立. 具体共定义了 4 种不同的微观话题主位推进模式见定义 10～定义 13. 图 2 给出了 4 种微观话题主位推进模式的可视化表示图.

定义 10. 放射型主位推进模式. 在一个篇章微观话题结构中，微观话题联接(MTLink)所连接的上一个端口是篇章基本话题中的主位，下一个端口连接的也是篇章基本话题中的主位，则构成放射型主位推进模式(MTS-based constant thematic progression, MTS-CosTP). 其形式化表示为

$$MTS-CosTP=(T_n,T_{n+1},V,\delta_n),$$

其中，$T_n \in T$，$T_{n+1} \in T$，T 为一个篇章中的篇章基本话题(EDTU)的主位(theme)集合；V 为连接成立的置信函数，$V(T_n,T_{n+1}) \in [0,1]$；$\delta_n \in \Gamma$，Γ 为同一个篇章中的微观话题联接(MTLink)的集合.

例 11. 两个绑匪(T_1)躲藏了起来(R_1)，他们($T_2=T_1$)绑住了小米的手脚(R_2).

例 11 中，第 2 句中的主位，即人称代词“他们”与上一句中的主位“两个绑匪”，存在指代关系，$T_2 \rightarrow T_1$，构成一个微观话题联接.

定义 11. 集中型主位推进模式. 在一个篇章微观话题结构中，微观话题联接(MTLink)所连接的上一个端口是篇章基本话题中的述位，下一个端口连接的也是篇章基本话题中的述位，则构成集中型主位推进模式(MTS-based centralized thematic progression, MTS-CenTP). 其形式化表示为

$$MTS-CenTP=(R_n,R_{n+1},V,\delta_n),$$

其中，$R_n \in R$，$R_{n+1} \in R$，R 为一个篇章中的篇章基本话题(EDTU)的述位(rheme)集合；V 为连接成立的置信函数，$V(R_n,R_{n+1}) \in [0,1]$；$\delta_n \in \Gamma$，Γ 为同一个篇章中的微观话题联接(MTLink)的集合.

例 12. 孩子们(T_1)笑了(R_1)，然后他们的母亲(T_2)也笑了($R_2=R_1$).

例 12 中，上下句述位包含“笑了”，存在重复关系，$R_2 \rightarrow R_1$，构成一个微观话题联接.

定义 12. 延续型主位推进模式. 在一个篇章微观话题结构中，微观话题联接(MTLink)所连接的上一个端口是篇章基本话题中的述位或述位的一部分，下一个端口连接的是篇章基本话题中的主位，则构成延续型主位推进模式(MTS-based simple linear thematic progression, MTS-SimTP). 其形式化表示为

$$MTS-SimTP=(R_n,T_{n+1},V,\delta_n),$$

其中，$R_n \in R$，$T_{n+1} \in T$，R 为一个篇章中的篇章基本话题(EDTU)的述位(rheme)集合；T 为同一个篇章中的篇章基本话题(EDTU)的主位(theme)集合；V 为连接成立的置信函数，$V(R_n,T_{n+1}) \in [0,1]$；$\delta_n \in \Gamma$，Γ 为同一个篇章中的微观话题联接(MTLink)的集合.

例 13. 我们的学校(T_1)是一个大花园(R_1)，花园里($T_2=R_1$)长满了各种花草(R_2).

例 13 中，后一句的主位中核心词“花园”，包含在前一句的述位中，$T_2 \rightarrow R_1$，构成一个微观话题联接.

定义 13. 交叉型主位推进模式. 在一个篇章微观话题结构中，微观话题联接(MTLink)所连接的上一个端口是篇章基本话题中的主位，下一个端口连接的是篇章基本话题中的述位或述位的一部分，则构成交叉型主位推进模式(MTS-based crossed thematic progression, MTS-CrsTP). 其形式化表示为

$$MTS-CrsTP=(T_n,R_{n+1},V,\delta_n),$$

其中,$T_n \in T, R_{n+1} \in R$,T 为一个篇章中的篇章基本话题(EDTU)的主位(theme)集合;R 为同一个篇章中的篇章基本话题(EDTU)的述位(rheme)集合;V 为连接成立的置信函数,$V(T_n, R_{n+1}) \in [0,1]$;$\delta_n \in \Gamma$,$\Gamma$ 为同一个篇章中的微观话题联接(MTLink)的集合.

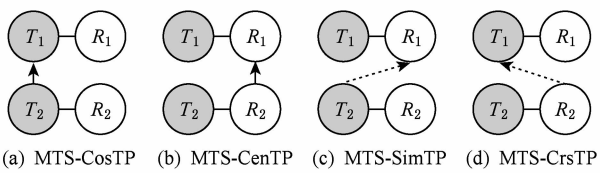


Fig. 2 Visual representation of four thematic progression patterns based on micro-topic scheme

图 2 4 种微观话题主位推进模式可视化表示图

例 14. 这只小猫(T_1)非常可爱(R_1),小朋友们(T_2)都非常喜欢它($R_2 = T_1$).

例 14 中,后一句的述位中核心词"它",与前一句主位"小猫"存在回指关系,$R_2 \rightarrow T_1$,构成一个微观话题联接.

在图 1 表示的例 1 汉语篇章话题结构中,前 4 个篇章微观话题联接,满足 MTS-CosTP 主位推进模式要求;最后第 5 个满足 MTS-CrsTP 主位推进模式要求.

3 篇章微观话题结构语料库(CDTC)

基于第 2 节微观话题结构形式化表示,为开展面向篇章理解的话题结构研究提供必要的语料资源,我们构建了汉语篇章话题结构语料库(Chinese discourse topic coupus).

CDTC 共包含 500 篇文档,其中原始自然句子(以句号或感叹号等结尾)共有 6 648 个.每个自然句子标注切分为多个篇章基本话题(EDTU),共得到 10 147 个篇章基本话题;每个篇章基本话题内部再次切分为主位和述位 2 部分.根据定义 2,由篇章基本话题(EDTU)及微观话题联接(MTLink)构建本语料库的微观话题结构(MTS),共标注 5 095 个 MTS;利用篇章话题结构内的微观话题联接(MTLink)构成 1 698 条篇章话题链(DTC);平均每个有效标注的篇章话题链连接 5.98 个 EDTU.

3.1 语料库构建

3.1.1 语料资源

为便于研究比较,我们构建 CDTC 的生语料资源来自宾州汉语树库 6.0 版(chtb0001-chtb0325, chtb0400~chtb0657).之所以采用上述 CTB 中的文档作为生语料,原因主要有 3 个:

1) 宾州汉语树库采用 PDTB 体系,自发布以来,在多类篇章结构分析任务中得到应用,具有较高的认可度;在该语料上完成微观话题结构的标注,有利于和其他已有研究开展比对.

2) 我们前期在结合 PDTB 和 RST 体系的基础上,提出了一种使用连接依存树的形式表示汉语篇章修辞结构的标注方案;并在此方案的指导下,已经选取上述中文树库 CTB6.0 上的 500 篇文档进行了篇章修辞结构的标注,这部分工作主要侧重为篇章连贯性研究提供语料资源.本文选用同样的 500 篇文档,采用微观话题结构标注体系,主要侧重为篇章衔接性研究提供语料资源,两者互为补充,为篇章话题结构提供联合研究资源.

3) 从应用角度考虑,篇章话题结构的研究,离不开组成篇章的字、词、句法等不同层次的特征.充分利用 CTB6.0 原有标注资源,结合我们的微观话题结构标注,为将来开展篇章有关的自然语言处理任务提供充分的语法、语义等多方面特征资源.

3.1.2 语料标注策略

总体指导原则是:一切从便于篇章理解的角度出发,制定相应的标注规范;采用计算机辅助人工半自动标注方法.根据主述位篇章微观话题结构及基于主位推进模式的微观话题联接机制,在一定规模的语料上试标注,针对包含照应、省略、替代等微观话题联接的标注及微观话题链识别提出具体的标注规范.标注规范注重可操作性,分别从判定原则、动态联接方法等方面入手制定,并给出例子详细说明,初步制定标注规范.进一步在较大规模语料上,实施和验证标注规范的科学性,适当做出调整,最终形成一套完整的汉语篇章微观话题结构标注规范.

针对不同阶段的标注对象,采用了 2 类具体的标注策略:用于篇章微观话题结构标注的自顶向下标注策略(top-down strategy)和用于篇章微观话题链标注的后向搜索标注策略(chain-backtracking strategy).

1) 自顶向下的语料标注策略.根据定义 2 可知,篇章微观话题结构的主要组成部分是篇章基本话题单元(EDTU),进一步包括篇章基本话题单元中的主位和述位.自顶向下的标注策略是指:对一段篇章内容,首先识别出全文中心主题,根据中心主题划分子主题所包含的主要的段落;之后,从每个段落中划分基本话题单元,层层递进,最后对其中包含的主位和述位进行切分.

在篇章微观话题结构的标注中使用自顶向下的策略,主要考虑:①由于篇章的话题结构呈现分层次特性,这种策略有利于自上而下在宏观上把握话题整体结构;②自顶向下的标注策略比较符合人类对篇章话题理解的一般心理过程,在汉语篇章连贯性分析和话题抽取研究中经常采用;③采用这种策略,相近子话题及相异子话题边界明晰,有利于提高篇章微观话题结构中篇章基本话题单元上下文之间的微观话题联接的标注准确率.很显然,微观话题联接更容易在相同或相近子话题内的篇章基本话题单元之间形成.这同时也有利于下一步篇章微观话题链的标注.

2) 后向搜索的语料标注策略.篇章微观话题链的标注是指多个微观话题联接形成连贯的链状结构.理论上来看,一条链的标注形成,可以分解为多个两两关联的微观话题联接的标注,这样一来,链标注的形成可以转化为微观话题联接的标注.对于微观话题联接,因为需要根据下文的主位或述位所表达出的话题含义,来回溯寻找上文所对应可能联接的主位或述位,因此,我们采用后向搜索的标注策略,即每次标注当前的主位或述位的微观话题联接,必定是回头看上文的对应主位或述位.所以,采用这种后向搜索标注策略,是由链结构本身特点所决定.

在采用这个策略进行标注的过程中,向后搜索几个篇章基本单元,即后向搜索的步数问题,是一个需要注意的关键问题.大多数情况下,标注者仅需要向上回溯一步即可,但也不排除某种特殊情形,需要回溯者向上搜索多步.如针对存在隐式主位或述位的情况下,有必要向后多步搜索.如例 15 所示.

例 15. (c)对此,[**浦东**][不是简单的采取…的做法,](d)[?][而是吸取…经验,](e)[?][聘请…专家,](f)[?][及时、迅速地制定和推出…文件,](g)[?][让这些经济活动…归入…处理流程.]

(c) In response to this,(**Pudong**) is not simply adopting an approach of "… "(d) (Instead, **Pudong**) is taking advantage of the experience of … (e) [by hiring appropriate domestic and foreign scholars and specialists], (f) [by actively and promptly formulating and issuing regulatory documents], (g) [and by ensuring that these economic activities are incorporated into …the legal system…]

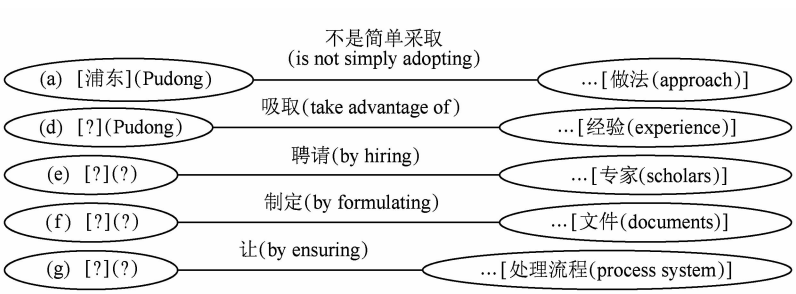


Fig. 3 Examples of multi-steps backward-searching for micro-topic linking

图 3 后向多步搜索微观话题联接的例子

图 3 描述了例 15 的主述位标注结构.图 3 中主位(d)只需向后回溯一步即可找到微观话题联接的对象"浦东";但是由于隐式主位(缺省主语)的存在,对于中文主位(e),(f),(g)要找到语义上的联接对象,就需要分别向上回溯 2 步、3 步、4 步,才能找到"浦东"这个联接对象了.英文平行语料的情况与中文相比略有不同,但也有类似缺省的情况.

3.1.3 人机结合的语料标注过程

CDTC 语料的标注工作主要由 5 位汉语言文学的本科生和 5 位计算机专业的本科生分成 2 组,在标注规范的指导下进行标注;本文作者和一位语言学专家指导核对,形成最终标准语料.标注分 4 阶段进行:

1) 为确保标注质量及其通用性,我们制定了初步的标注规范并用来培训标注者,同时完成了计算机辅助标注工具的开发;

2) 为确保标注一致,所有参与的标注人员首先分别标注相同的 50 篇文档,然后集中逐一校对讨论,讨论内容包括篇章基本话题单元、微观话题结构识别及其构成的话题链识别等在内的所有标注内容,统一重新修订完成新的标注规范;

3) 标注人员分组完成语料的标注,这个数据用来计算语料标注的一致性;制定标注规范和开展标注实践,必须反复迭代进行,多轮完善后才能得到较为合适的标注规范;

4) 根据最终的标注规范,逐一校对标注语料,最终合并形成可发布的汉语篇章话题结构语料库 CDTC.

在CDTC语料库标注时,首先导入我们前期已经完成的篇章修辞结构标注处理的语料,作为需要话题结构标注的生语料,然后利用计算机辅助工具生成语料的可视化篇章结构,以辅助人工分析话题结构;通过人工分析识别主述位,寻找候选主述位,建立话题链接关系.为评估多人标注完成的语料是否达到一致性要求,我们利用一致性检验方法完成了相应的一致性计算,并统计分析了所完成的标注语料结果.此外,为了克服手工标注生文本费时费力,且容易出错的问题,我们设计开发了汉语篇章微观话题结构计算机辅助标注系统,如图4所示,功能模块包含有篇章结构预处理、计算机辅助可视化结构生成、语料半自动标注、标注结果生成、语料自动统计和一致性自动计算等.其中在核心功能语料半自动标注模块中,还细分为微观话题结构中主述位标注、微观话题链识别标注等操作.

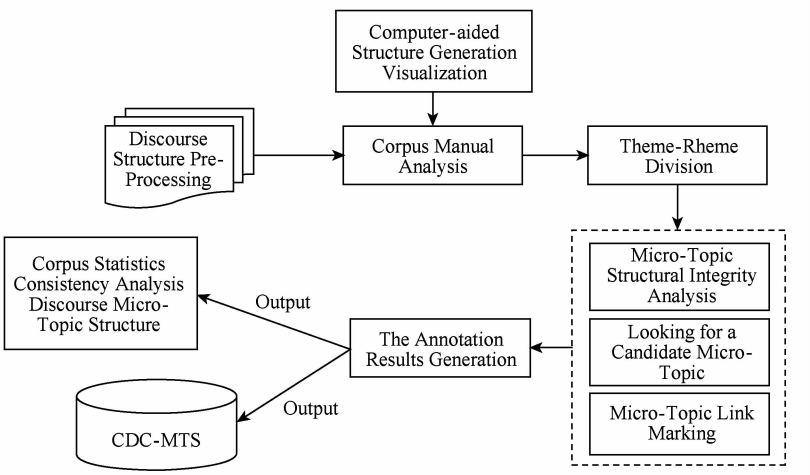


Fig. 4 Processing flow of annotation platform for Chinese discourse micro-topic scheme

图4 汉语篇章微观话题结构标注平台处理流程

3.2 标注规范

CDTC语料库定义了独立的标注规范,由标注者根据自己对语料的理解进行标注,标注内容包括篇章基本话题单元和篇章微观话题结构.篇章基本话题的标注较为简单,我们要求标注者首先确认包含且仅包含一个谓词的独立子句,然后根据该子句是否为最小的独立表达意图性的单位来确认其是否为篇章基本话题.然而,由于篇章微观话题结构标注需要涉及篇章基本话题中的主位和述位识别、上下篇章基本话题之间的微观话题联接及其衔接类型识别,因此标注难度极大,为此我们为标注者制订了一整套标注规范,主要是针对篇章微观话题结构的标注进行说明.

3.2.1 标注总则

首先我们通过一个具体的例子来分析我们标注方案中标注的篇章微观话题结构具体内容.

例16. 1)[[浦东]$_{\text{Satellite}(T_1)}$开发开放]$_{T_1}$[是一项…工程,]$_{R_1}$ 2)[[null]$_{\text{Satellite}(T_2)}$(因此)大量面对的]$_{\langle \text{Satellite}(T_2)=T_1\rangle}$[是…新状况、新事务.]$_{R_2}$ 3)[(对此),[浦东]$_{\langle T3=\text{Satellite}(T_2)\rangle}$]$_{T_3}$[不是简单的采取…的做法,]$_{R_3}$ 4)[null]$_{\langle T_4=T_3\rangle}$[而是吸取…发展经验,]$_{R_4}$ 5)[null]$_{\langle T_5=T_4\rangle}$[聘请…专家,]$_{R_5}$ 6)[null]$_{\langle T_6=T_5\rangle}$[及时、迅速地制定和推出…文件,]$_{R_6}$ 7)[null]$_{\langle T_7=R_6\rangle}$[让这些经济活动…被归入合法的处理流程.]$_{R_7}$

例16所示篇章微观话题结构采用图形化表示的例子如图5所示,其中例16中字母所标记的语段表示篇章基本话题(EDTU),T_n前面的语段表示主位,T_n后面的语段表示述位,用R_n表示;各篇章基本话题通过连接主述位的微观话题联接组合后形成微观话题结构,进而通过再组合形成更高一级篇章话题结构(其组合过程也是微观话题联接构建微观话题链的过程);如此层层组合,最后形成中心篇章话题结构,并且形式上表现为微观话题链.从图5可知,例16所示篇章最后可以由2条虚线构成的箭头作为链的头结点表示(由图5中指向中心圆的2条虚线构成的箭头作为链的头结点表示),并形成整个篇章的核心话题.

3.2.2 篇章微观话题结构标注

结合主述位理论、RST、PDTB、汉语复句理论、汉语句群理论和广义话题结构理论等的研究,我们提出用主述位构建微观话题链的形式表示汉语的篇章话题结构,主要针对的是篇章微观话题结构的标注,其标注方案如下.

〈EDTU ID=[1..N]〉

<MTS ID=[1..N] /*ID号*/

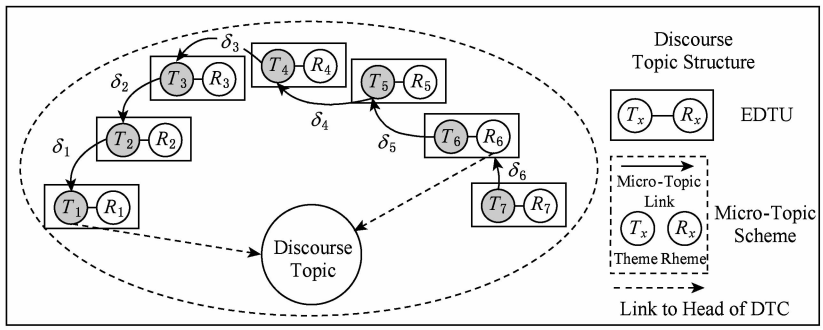


Fig. 5 The Instance of Micro-Topic Scheme for Example 16

图 5 例 16 的微观话题结构实例图

TYPE=[Entity|Event] /* 实体、事件 */

POSITION=[Theme|Rheme] /* 主位、述位 */

LOCATION=[Root|NotR] /* 初次出现、非初次出现 */

KEY=[Complex|Satellite|Nucleus] /* 组合、辅助、核心 */

RTYPE=[NotZ|Zero] /* 非零主述位、零主述位 */

LINKID=[0..N] /* 微观话题联接的上级 ID 号 */

LINKTYPE=[***] /* 衔接的类型 */

USETIME=[Numbers] /* 标注用时统计,体现标注评定难度,单位:秒 */

>主位或述位对象〈/MTS〉……

〈/EDTU〉

基于上述篇章微观话题结构的标注方案,我们标注篇章基本话题、篇章微观话题结构,其中篇章微观话题结构包含联合主/述位,核心主/述位,篇章微观话题联接及其联接关系类型,相关标注说明如下.

EDTU 中的 ID 表示当前标注对象在当前文档中的唯一标识号,起始号为 1,增幅为 1,逐步递增.这里的标注对象是指当前篇章中的 EDTU.

例 17 表示的即为一个独立 EDTU,其中包含且仅包含一个"是"作为谓词结构.

例 17. 〈EDTU ID="1"〉浦东开放建设是一项…工程〈/EDTU〉.

MTS 中的 ID 表示当前标注对象在当前文档中的唯一标识号,起始号为 1,增幅为 1,逐步递增.这里的标注对象是指包含在当前 EDTU 中的主位或述位.

例 17 中的"浦东开放建设"识别为主位,用 T_1 表示;剩余部分"是一项…工程"识别为述位,用 R_1 表示,如例 18 中 1)所示;采用 XML 标记对正式标注形成例 18 中 2).

例 18. 1) [浦东开放建设]$_{T_1}$[是一项…工程]$_{R_1}$.

2) 〈MTS ID="1"…〉浦东开放建设〈/MTS〉〈MTS ID="2"…〉是一项…工程,〈/MTS〉

MTS 中的 TYPE 表示当前标注对象的类型,共有 2 类取值,分别为"Entity","Event",分别表示"实体"、"事件".这里的标注对象是指包含在当前 EDTU 中的主位或述位.有关实体、事件的定义,我们采用 PDTB 体系的同类概念.

例 19 所示即为"实体"的标注类型.其中的"上海浦东"识别为主位,TYPE 值为"Entity",即表示"实体"类型;剩余部分"近年来颁布实行了…文件"识别为述位.从词性角度来看,标注为实体的主位或述位一般是名词或名词短语.

例 19. 〈MTS ID="1" TYPE="Entity"…〉上海浦东〈/MTS〉近年来颁布实行了…文件.

从词性角度来看,标注为事件的主位或述位一般是动词、动+宾、动词短语等.例 20 中 1)所示即为标注为"事件"类型的述位结构,其中的"动+宾"结构,"…采取…做法"识别为述位,TYPE 值为"Event",即表示"事件"类型.而例 20 中 2)所示即为标注为"事件"类型的主位结构,其中"浦东开放建设"核心词是"开放建设",从词性角度来看,属于动宾短语结构.

例 20. 1) …,浦东〈MTS ID="6" TYPE="Event"…〉不是简单的采取…做法,〈/MTS〉

2) 〈MTS ID="1" TYPE="Event"…〉浦东开放建设〈/MTS〉是一项…工程

MTS 中的 POSITION 表示当前标注对象的位置类型,共有 2 类取值,分别为"Theme"、"Rheme",分别表示"主位"、"述位".

例 21 所示 POSITION 取值"Theme",即表示带有"主位"标注类型的标注语料.

例 21. 〈MTS ID="1" TYPE="Entity" POSITION="Theme"…〉苏州经济建设〈/MTS〉取得可喜成果.

例 22 所示 POSITION 取值"Rheme",即表示带有"述位"标注类型的标注语料.

例 22. 西藏金融工作〈MTS ID="1" TYPE="Event" POSITION="Rheme"…〉取得显著成绩.〈/MTS〉

MTS 中的 LOCATION 表示当前标注对象是否初次出现,共有 2 类取值,分别为"Root"、"NotR",分别表示"初次出现"、"非初次出现".

例 23 表示 2 个相邻 EDTU,分别以例 23 中 1)和例 23 中 2)为编号.其中例 23 中 1)所标注对象"世界上最大的…国际承包商"的属性 LOCATION 取值为"Root",即表示该对象为初次标注;后续例 23 中 2)能够与之形成相关联的微观话题链,则例 23 中 1)所含标注对象为链首结点.例 23 中 2)所标注对象"其中不少公司"与上文例 23 中 1)所标注对象存在关联关系,所以认定当前例 23 中 2)所标注对象不是首次出现,其属性 LOCATION 取值为"NotR".

例 23. 1)〈EDTU ID="1"〉〈MTS ID="18" TYPE="Entity" POSITION="Theme" LOCATION="Root"…〉世界上最大的…国际承包商〈/MTS〉已进入中国,〈/EDTU〉

2)〈EDTU ID="2"〉〈MTS ID="19" TYPE="Entity" POSITION="Theme" LOCATION="NotR"…〉其中不少公司〈/MTS〉与中国公司合资合作进行建设.〈/EDTU〉

MTS 中的 KEY 表示当前标注对象在意图表达上的重要程度,共有 3 类取值,分别为"Complex","Nucleus","Satellite",分别表示"组合标注"、"核心标注"、"辅助标注".这里的标注对象是指包含在当前 EDTU 中的"主位"或"述位".

组合标注是当前标注属性 KEY 的默认取值.组合标注内部可以包含核心标注和辅助标注,也可以不包含其他任何标注;但是核心标注或辅助标注必须包含在组合标注内部.

核心标注所标注的对象,能够体现外围组合标注所标注对象的意图,是其核心语义的体现.辅助标注所标注的对象,是核心标注所标注对象的辅助成分,不是外围组合标注对象的意图.

例 24 所示即为不包含其他 2 个标注的组合标注类型;例 25 和例 26 分别表示含有核心标注和辅助标注的组合标注类型.

例 24. 〈MTS ID="1" TYPE="Entity" POSITION="Theme" LOCATION="Root" KEY="Complex"…〉中国三资企业人民币贷款余额〈/MTS〉近一千亿元.

例 25 所示即为包含核心标注的组合标注类型,其中 ID 为 1.1 的标注对象"中国进出口银行"即为 ID 为 1 的标注对象"去年十月,中国进出口银行"的核心语义体现,换句话说,在标注 ID 为 1 的这个标注对象中,作者主要的意图是表达"中国进出口银行"这个实体.

例 25. 〈MTS ID="1" TYPE="Entity" POSITION="Theme" LOCATION="Root" KEY="Complex"…〉去年十月,〈MTS ID="1.1" TYPE="Entity" POSITION="Theme" LOCATION="Root" KEY="Nucleus"…〉中国进出口银行〈/MTS〉〈/MTS〉聘请日本野村证券公司作顾问,

例 26 所示即为包含辅助标注的组合标注类型,其中 ID 为 18.2 的标注对象"目前,"即为核心标注"外商投资企业"的辅助,表示时间状态.

例 26. 〈MTS ID="18" TYPE="Entity" POSITION="Theme" LOCATION="Root" KEY="Complex"…〉〈MTS ID="18.2" TYPE="Entity" POSITION="Theme" LOCATION="Root" KEY="Satellite"…〉目前,〈/MTS〉约有十五万家〈MTS ID="18.1" TYPE="Entity" POSITION="Theme" LOCATION="NotR" KEY="Nucleus"…〉外商投资企业〈/MTS〉〈/MTS〉在中国银行开立帐户.

MTS 中的 RTYPE 表示当前标注对象是否属于缺省,共有 2 类取值:"NotZ","Zero",分别表示"非零结构"或"零结构".这里的标注对象是指包含在当前 EDTU 中的"主位"或"述位".我们把零结构这种情况也称为隐式主位或隐式述位现象;对应非零结构则称为显式主位或显式述位.

从对当前 EDTU 完整理解的角度来看,"零结构"表示当前标注对象处于缺省或省略状态,而缺省或省略的内容,可以从上文中关联得到.例 27 表示了这种关联情况,其中例 27 中 1)所示为"非零结构"的类型,而例 27 中 2)则表示了"零结构"的类型,两者是上下文关联的.

例 27. 1)〈EDTU ID="1"〉〈MTS ID="1" TYPE="Event" POSITION="Theme" LOCATION="Root" KEY="Complex" RTYPE=

"NotZ"…〉…外商投资项目〈/MTS〉近二十五点九万个,〈/EDTU〉

2)〈EDTU ID="2"〉〈MTS ID="3" TYPE="Event" POSITION="Theme" LOCATION="NotR" KEY="Complex" RTYPE="Zero"…〉null〈/MTS〉实际利用外资…美元.〈/EDTU〉

和"非零结构"相反,不存在显式内容的标注对象即为"零结构",如例 27 中 2)所示.其中,零结构所标注对象不存在,我们用"null"来代替该标注对象.此外,零结构和上文非零结构的关联关系,我们也有标注体现,主要采用 MTS 中的 LINKID 和 LINKTYPE 两个属性来表示,相关内容详见第 3.2.4 节有关话题链的特殊标注规则.

MTS 中的 USETIME 表示标注者在识别当前标注对象过程中思考分析所用的时间,反映标注对象的语义理解难度,它由标注辅助程序自动计算得到.这里的标注对象是指包含在当前 EDTU 中的主位或述位.

USETIME 的计算方法是由标注辅助程序自动计算从上一个标注结束开始,到当前这个标注结束时的间隔时间,单位为 s,例 28 所示其中标注时间 USETIME 为 15 s.

例 28. 〈MTS ID="1" TYPE="Entity" POSITION="Theme" LOCATION="NotR" KEY="Complex" RTYPE="NotZ" USETIME="15"…〉苏州海关驻张家港办事处〈/MTS〉于日前成立.

3.2.3 篇章微观话题链标注

MTS 中的 LINKID 和 LINKTYPE 是一对相关属性.

MTS 中的 LINKID 表示与当前标注对象存在微观话题联接(micro-topic link)关系的上文标注对象的 ID 号.这里的标注对象是指包含在当前 EDTU 中的主位或述位.LINKID 属性的取值受到上文 MTS 中 ID 的约束,也就是说,上文出现的 ID 属性值是当前 LINKID 属性的取值范围.但是也有特殊情况,就是当前标注对象是首次出现,则默认当前 LINKID 属性取值为"0",即 MTS 中的 LOCATION 属性取值为"Root"时,如例 29 中 1)所示.例 29 中 2)表示了正常情况下的 LINKID 取值.

例 29 中 1)标注的主位内容是"…保税区",该内容在全文中首次出现,所以 LOCATION 取值为"Root",对应 LINKID 取值为"0".

例 29. 1)〈EDTU ID="1"〉〈MTS ID="1"… USETIME="10" LINKID="0" LINKTYPE="Empty"〉…保税区〈/MTS〉今后五年将充分发挥…优势,〈/EDTU〉

2)〈EDTU ID="2"〉〈MTS ID="3"… USETIME="17" LINKID="1"…〉null〈/MTS〉以高新技术产业为先导,〈/EDTU〉.

例 29 中 2)和例 29 中 1)是上下文关系.例 29 中 2)中标注的主位是缺省主位,属性 LINKID 取值为"1",即代表例 29 中 1)中的主位标记属性 ID 的值,其指向例 29 中 1)中的主位标注内容,即"…保税区".如果将例 29 中 1)的主位标注内容填入例 29 中 2),即可构成完整的一个篇章基本话题(EDTU):"…保税区以高新技术产业为先导,".

属性 LINKID 关联了上下文 EDTU,其实质作用是实现了篇章之间的衔接(Cohesion)关系.

MTS 中的 LINKTYPE 表示当前标注对象与上文标注对象之间存在的话题联接关系的类型,分为"照应"、"省略"、"替代"、"重复"、"同义"、"反义"、"具体抽象化"、"抽象具体化"、"整体局部化"、"局部整体化"、"搭配"共 11 种类型,分别取值为"Reference","Ellipsis","Substitution","Repetition","Synonym","Antisense","Abstraction","Instantiation","Partialization","Integration","Collocation"(详细定义见第 2 节).此外,特殊情况下,当 LINKID 取值为"0"时,LINKTYPE 取值为"Empty",表示不存在任何话题联接关系,如上文例 29 中 1)中的 LINKTYPE 取值即属于这种特殊情况.这里的标注对象是指包含在当前 EDTU 中的主位或述位.

当存在正常话题联接时,话题联接关系如下文例 30～例 39 所示.

1) 照应.例 30 中 1)和例 30 中 2)属于同一篇章中的上下文.例 30 中 1)中标注的述位是一个包含核心标注(nucleus)的组合标注(complex)述位,其中下划线划出的内容就是核心标注的内容,即"九百九十五点六亿元";该核心标注内容与下文例 30 中 2)中的主位标注"这一数字"形成照应关系.从例子 30 中 2)中对这个主位标注的属性 LINKID 和 LINKTYPE 取值可以看到,LINKID 的取值"2.1"即等于例 30 中 1)所标注照应语(下划线标出部分)所在 MTS 标注的 ID 取值;而 LINKTYPE 的取值"Reference"即表示当前话题联接关系的类型为"照应".

例 30. 1)〈EDTU ID="1"〉〈MTS ID="1"… LINKTYPE="Empty"〉…人民币贷款余额〈/MTS〉

〈MTS ID="2"…LINKTYPE="Empty"〉已达〈MTS ID="2.1"…LINKID="0" LINKTYPE="Empty"〉九百九十五点六亿元〈/MTS〉〈/MTS〉,〈/EDTU〉

2)〈EDTU ID="2"〉〈MTS ID="3"…LINKID="2.1" LINKTYPE="Reference"〉这一数字〈/MTS〉比上年末增加二百零三点三亿元,〈/EDTU〉.

2)省略.例31中1)和例31中2)属于同一篇章中的上下文.例31中1)中标注的主位是一个包含核心标注(nucleus)的组合标注(complex)主位,其中下划线划出的内容就是核心标注的内容,即"去年苏州台资企业缴纳的所得税";该核心标注内容与下文例31中2)中的主位空标注"null"形成省略关系.从例子31中2)中对这个主位标注的属性LINKID和LINKTYPE取值可以看到,LINKID的取值"9.1"即等于例31中1)所标注照应语(下划线标出部分)所在MTS标注的ID取值;而LINKTYPE的取值"Ellipsis"即表示当前话题联接关系的类型为"省略".

例31. 1)〈EDTU ID="5"〉〈MTS ID="9"…LINKTYPE="Empty"〉据…统计,〈MTS ID="9.1"…LINKID="0" LINKTYPE="Empty"〉去年苏州台资企业缴纳的所得税〈/MTS〉〈/MTS〉〈MTS ID="10"…LINKTYPE="Empty"〉达三点六七亿元,〈/MTS〉〈/EDTU〉

2)〈EDTU ID="6"〉〈MTS ID="11"…LINKID="9.1" LINKTYPE="Ellipsis"〉null〈/MTS〉比上年增长百分之五十七点一.〈/EDTU〉

3)替代.例32中1)和例32中2)属于同一篇章中的上下文,但并非直接相邻上下文.例32中1)中标注的主位是一个包含核心标注(nucleus)和辅助标注(satellite)的组合标注(complex)主位,其中下划线划出的内容就是辅助标注的内容,即"西藏";该辅助标注内容与下文例32中2)中的主位标注"全区"形成替代关系.从例子32中2)中对这个主位标注的属性LINKID和LINKTYPE取值可以看到,LINKID的取值"1.1.1"即等于例32中1)所标注辅助标注内容(下划线标出部分)所在MTS标注的ID取值;而LINKTYPE的取值"Substitution"即表示当前话题联接关系的类型为"替代".

例32. 1)〈EDTU ID="1"〉〈MTS ID="1"…LINKTYPE="Empty"〉…期间,〈MTS ID="1.1"…LINKTYPE="Empty"〉〈MTS ID="1.1.1"…LINKID="0" LINKTYPE="Empty"〉西藏〈/MTS〉金融〈/MTS〉体制改革〈/MTS〉坚持…方针,〈/EDTU〉…

2)〈EDTU ID="4"〉〈MTS ID="7"…LINKTYPE="Empty"〉去年,〈MTS ID="7.1"…LINKID="1.1.1" LINKTYPE="Substitution"〉全区〈/MTS〉各项存款〈/MTS〉首次突破…大关.〈/EDTU〉

4)重复.例33中1)和例33中2)属于同一篇章中的上下文.例33中1)中标注的主位是一个包含核心标注(nucleus)的组合标注(complex)主位,其中下划线划出的内容就是核心标注的内容,即"韩国";该核心标注内容与下文例33中2)中的述位辅助标注"韩国"(下划线划出内容)形成重复关系.从例子33中2)中对这个述位标注的属性LINKID和LINKTYPE取值可以看到,LINKID的取值"3.1"即等于例33中1)所标注内容(下划线标出部分)所在MTS标注的ID取值;而LINKTYPE的取值"Repetition"即表示当前话题联接关系的类型为"重复".

例33. 1)〈EDTU ID="2"〉〈MTS ID="3"…LINKTYPE="Empty"〉截止…,〈MTS ID="3.1"…USETIME="8"LINKID="0" LINKTYPE="Empty"〉韩国〈/MTS〉〈/MTS〉在华投资企业总数为…家,〈/EDTU〉

2)〈EDTU ID="3"〉〈MTS ID="5"…LINKTYPE="Empty"〉中国〈/MTS〉〈MTS ID="6"…LINKTYPE="Empty"〉已成为〈MTS ID="6.1"…USETIME="14" LINKID="3.1" LINKTYPE="Repetition"〉韩国〈/MTS〉最大的投资对象国.〈/MTS〉〈/EDTU〉

5)同义.例34中1)和例34中2)属于同一篇章中的上下文.例34中1)中标注的述位是一个组合标注(complex)述位,其中下划线划出的内容就是标注的内容,即"已投入使用,";该标注内容与下文例34中2)中的述位"运转正常,"(下划线划出内容)形成同义关系.从例子34中2)中对这个述位标注的属性LINKID和LINKTYPE取值可以看到,LINKID的取值"12"即等于例34中1)所标注内容(下划线标出部分)所在MTS标注的ID取值;而LINKTYPE的取值"Synonym"即表示当前话题联接关系的类型为"同义".

例34. 1)〈EDTU ID="6"〉〈MTS ID="11"…LINKTYPE="Empty"〉二千门程控电话〈/MTS〉

〈MTS ID="12"…LINKID="0" LINKTYPE="Empty"〉已投入使用,〈/MTS〉〈/EDTU〉

2) 〈EDTU ID="7"〉〈MTS ID="13"…LINKTYPE="Empty"〉十千伏高压电路〈/MTS〉〈MTS ID="14"…LINKID="12" LINKTYPE="Synonym"〉运转正常,〈/MTS〉〈/EDTU〉

6) 反义. 例 35 中 1)和例 35 中 2)属于同一篇章中的上下文. 例 35 中 1)中标注的述位是一个包含核心标注(nucleus)的组合标注(complex)主位,其中下划线划出的内容就是核心标注的内容,即"一个默默无闻的小渔村";该核心标注内容与下文例 35 中 2)中的述位核心标注"一个现代化都市的框架"(下划线划出内容)形成反义关系,或者说是一种对比反差关系. 从例子 35 中 2)中对这个述位标注的属性 LINKID 和 LINKTYPE 取值可以看到,LINKID 的取值"2.1"即等于例 35 中 1)所标注内容(下划线标出部分)所在 MTS 标注的 ID 取值;而 LINKTYPE 的取值"Antisense"即表示当前话题联接关系的类型为"反义".

例 35. 1) 〈EDTU ID="1"〉〈MTS ID="1"…LINKTYPE="Empty"〉数年前,北海〈/MTS〉〈MTS ID="2"…LINKTYPE="Empty"〉还是北部湾〈MTS ID="2.1"…LINKTYPE="Empty"〉一个默默无闻的小渔村〈/MTS〉,〈/MTS〉〈/EDTU〉

2) 〈EDTU ID="2"〉〈MTS ID="3"…USETIME="12" LINKID="1" LINKTYPE="Repetition"〉然而三五年时间北海〈/MTS〉〈MTS ID="3"…USETIME="12" LINKID="2" LINKTYPE="Antisense"〉已建成了〈MTS ID="3"…USETIME="12" LINKID="2.1" LINKTYPE="Antisense"〉一个现代化都市的框架〈/MTS〉,〈/MTS〉〈/EDTU〉

7) 具体抽象化. 例 36 中 1)和例 36 中 2)属于同一篇章中的上下文. 例 36 中 1)中标注的主位是一个包含辅助标注(satellite)的组合标注(complex)主位,其中下划线划出的内容就是辅助标注的内容,即"齐宝芳";该标注内容与下文例 36 中 2)中的主位组合标注"韩国"(下划线划出内容)形成重复关系. 从例子 36 中 2)中对这个述位标注的属性 LINKID 和 LINKTYPE 取值可以看到,LINKID 的取值"3.1"即等于例 36 中 1)所标注内容(下划线标出部分)所在 MTS 标注的 ID 取值;而 LINKTYPE 的取值"Abstraction"即表示当前话题联接关系的类型为"具体抽象化".

例 36. 1) 〈EDTU ID="2"〉〈MTS ID="3"…LINKTYPE="Empty"〉〈MTS ID="3.1"…LINKTYPE="Empty"〉齐宝芳〈/MTS〉个人投资〈/MTS〉共计….〈/EDTU〉

2) 〈EDTU ID="3"〉〈MTS ID="5"…USETIME="8" LINKID="3.1" LINKTYPE="Abstraction"〉像齐家这样…的农民〈/MTS〉在当地并不在少数.〈/EDTU〉

8) 抽象具体化. 例 37 中 1)和例 37 中 2)属于同一篇章中的上下文. 例 37 中 1)中标注的主位是一个组合标注(complex)主位,其中下划线划出的内容就是组合标注的内容,即"一个主营电信设备的民营科技企业";该组合标注内容与下文例 37 中 2)中的主位辅助标注"深圳华为技术有限公司"(下划线划出内容)形成一种抽象与具体的关系,即"一个…民营科技企业"是个抽象的概念,而"深圳华为技术有限公司"是其具体的一个实例. 从例子 37 中 2)中对这个主位标注的属性 LINKID 和 LINKTYPE 取值可以看到,LINKID 的取值"3"即等于例 37 中 1)所标注内容(下划线标出部分)所在 MTS 标注的 ID 取值;而 LINKTYPE 的取值"Instantiation"即表示当前话题联接关系的类型为"抽象具体化".

例 37. 1) 〈EDTU ID="2"〉〈MTS ID="3"…LINKID="0"…〉一个…民营科技企业〈/MTS〉创造了…发展速度.〈/EDTU〉

2) 〈EDTU ID="3"〉〈MTS ID="5"…LINKID="0"…〉〈MTS ID="5.1"…USETIME="10"LINKID="3" LINKTYPE="Instantiation"〉深圳华为技术有限公司〈/MTS〉今年销售收入〈/MTS〉达一百亿元人民币,

9) 整体局部化. 例 38 中 1)和例 38 中 2)属于同一篇章中的上下文. 例 38 中 1)中标注的是一个组合标注(complex)主位,其中下划线划出的内容就是标注的内容,即"地处长江中游的湖南省";该标注内容与下文例 38 中 2)中的主位组合标注"境内湘江、资江、沅江和澧水"(下划线划出内容)形成整体与局部的关系. 因为从地理区划角度,"湘江、资江、沅江和澧水"都是"湖南省"的一部分. 从例子 38 中 2)中对这个主位标注的属性 LINKID 和 LINKTYPE 取值可以看到,LINKID 的取值"1"即等于例 38 中 1)所标注内容(下划线标出部分)所在 MTS 标注的 ID 取值;而 LINKTYPE 的取值"Partialization"即表示当前话题联接关系的类型为"整体局部化".

例 38. 1)〈EDTU ID="1"〉〈MTS ID="1"…USETIME="8"…〉地处长江中游的湖南省〈/MTS〉,是中国江河湖泊水系最复杂的省份之一,〈/EDTU〉…

2)〈EDTU ID="3"〉〈MTS ID="5"…USETIME="8" LINKID="1" LINKTYPE="Partialization"〉境内湘江、资江、沅江和澧水〈/MTS〉经洞庭湖流入长江.〈/EDTU〉

10) 局部整体化. 还可以分成"一对一"整体化和"多对一"整体化. "一对一"整体化是指构成联接关系的上下文都是单独一个标注对象;而"多对一"整体化,则指构成联接关系的上文有多个标注对象,而下文只有一个标注对象. "多对一"整体化情况比较特殊,我们在 3.3.4 节详细说明;"一对一"整体化如下例 39 所示.

11) "一对一"整体化. 例 39 中 1)和例 39 中 2)属于同一篇章中的上下文. 例 39 中 1)中标注的主位是一个包含辅助标注(satellite)的组合标注(complex)主位,其中下划线划出的内容就是组合标注的内容,即"湄洲湾开发区";该辅助标注内容与下文例 39 中 2)中的主位标注"湄州湾"(下划线划出内容)形成一种局部和整体的关系,即"湄洲湾开发区"是"湄洲湾"的一部分,"湄洲湾"是"湄洲湾开发区"的整体表述. 从例子 39 中 2)中对这个主位标注的属性 LINKID 和 LINKTYPE 取值可以看到,LINKID 的取值"11.1"即等于例 39 中 1)所标注内容(下划线标出部分)所在 MTS 标注的 ID 取值;而 LINKTYPE 的取值"Integration"即表示当前话题联接关系的类型为"局部整体化".

例 39. 1)〈EDTU ID="6"〉〈MTS ID="11"…LINKID="0"…〉届时,〈MTS ID="11.1"…LINKID="0" LINKTYPE="Empty"〉湄州湾开发区〈/MTS〉的工业产值〈/MTS〉与目前福建全省的经济实力基本相当.〈/EDTU〉

2)〈EDTU ID="7"〉〈MTS ID="13"…LINKID="11.1" LINKTYPE="Integration"〉湄州湾〈/MTS〉位于…的中点,〈/EDTU〉

3.2.4 特殊标注规则

3.2.2 和 3.2.3 节分别介绍了篇章微观话题结构和篇章微观话题链标注的基本规则,其标注的 2 个对象之间的关系都是"一对一"的关系. 然而,在某些特殊情况下,后续对象语义的具体内容会包含前续多个标注对象,形成"多对一"的包含关系. 本节介绍这种情况,如例 40 所示.

例 40. 1)~4)属于同一篇章中的上下文.

1) 监督体系不健全.

2) 产品质量低劣,

3) 假冒伪劣屡禁不绝,

4) 对中国经济和社会发展造成严重危害.

例 40 中 1)~4)分别表达 3 个篇章基本话题(EDTU),而在行文到 4)时,采用了缺省表示的方式,缺省的内容正是上文 1)~3)所表达的话题. 如果补充完整,可以加上个缺省代词"这",即第 4)句话可以补充为"[这]对…造成严重危害." 此时,这里的"这"指代前面 3 个基本话题的合并内容,而不仅仅是单个基本话题. 我们把这种情况称为微观话题的联合指代现象.

联合指代现象在我们的标注规则中,主要涉及 2 个标注标记:1)在 MTS 中的属性 LINKTYPE 的取值需要为"Integration";2)在 MTS 中的属性 LINKID 需要包含多个取值,并利用这些取值与上文篇章基本话题建立关联关系.

以例 40 所示内容开展标注形成例 41 所示,即为"联合指代"的标注类型. 例 41 中 1)中的划线内容"监督体系"识别为主位,MTS ID 取值为"23";例 41 中 2)中的划线内容"产品质量"识别为主位,MTS ID 取值为"25";例 41 中 3)中的划线内容"假冒伪劣"识别为主位,MTS ID 取值为"27". 上述 3 个标注对象联合表示,成为下文例 41 中 4)中的缺省主位. 而在例 41 中 4)中,则通过 LINKID 取多个值"23,25,27"将缺省主位与上文的标注对象形成关联关系.

例 41. 1)〈EDTU ID="12"〉〈MTS ID="23"…USETIME="13" LINKID="0" LINKTYPE="Empty"〉监督体系〈/MTS〉不健全.〈/EDTU〉

2)〈EDTU ID="13"〉〈MTS ID="25"…USETIME="15" LINKID="0" LINKTYPE="Empty"〉产品质量〈/MTS〉低劣,〈/EDTU〉

3)〈EDTU ID="14"〉〈MTS ID="27"…USETIME="8" LINKID="0" LINKTYPE="Empty"〉假冒伪劣〈/MTS〉屡禁不绝,〈/EDTU〉

4)〈EDTU ID="15"〉〈MTS ID="29"…USETIME="31" LINKID="23, 25, 27" LINKTYPE="Integration"〉null〈/MTS〉对…造成严重危害.〈/EDTU〉

3.3 语料库统计与分析

目前 CDTC 共有 500 个文档(chtb001-chtb0657),原始自然句子(以句号或感叹号等结尾)共有 6 648

个.每个自然句子标注切分为多个篇章基本话题(EDTU);每个篇章基本话题内部再次切分为2部分,即组合主位(complex theme)和组合述位(complex rheme).根据定义2,由篇章基本话题(EDTU)及微观话题联接(MTLink)构建本语料库的微观话题结构(MTS),共标注5 095个MTS;由MTS再次通过MTLink递归构建形成1 698篇章话题链,平均每个有效标注的篇章话题链连接5.98个EDTU.

下面分别从篇章基本话题(EDTU)、篇章基本话题中的主位(theme)和述位(rheme)、微观话题联接(MTLink)、微观话题结构(MTS)、篇章话题结构(DTS)等方面对CDTC语料库进行详细的统计分析.

3.3.1 语料库统计

3.3.1.1 篇章基本话题(EDTU)统计

篇章基本话题是本语料库的基础单元,共有效标注10 147个篇章基本话题;共有500个自然篇章文本,每个文本平均包含约20.3个篇章基本话题.表1是对篇章基本话题的统计.

Table 1 Distribution of the Elemental Discourse Topic Unit

表1 篇章基本话题在篇章中的分布统计

Number of EDTUs Each Unit	Unit=Sentence		Unit=Paragraph	
	Number	Number/Total/%	Number	Number/Total/%
1	2 979	53.28	900	28.54
2	1 400	25.04	634	20.11
3	679	12.14	525	16.65
4	277	4.95	343	10.88
5	123	2.20	256	8.12
6	55	0.98	175	5.55
7	53	0.95	117	3.71
8	11	0.20	62	1.97
9	4	0.07	45	1.43
>9	10	0.18	96	3.04
Total	5 591	100	3 153	100

其中,包含EDTU的自然句子共有5 591个(另有1 057个句子仅包含篇章首部的作者信息、篇章尾部的固定词等,故排除在外),3 153个自然段落.

3.3.1.2 主位-述位(Theme-Rheme)统计分析

CDTC语料库中每个篇章基本话题都一分为二,分别形成10 147个独立的组合主位(complex theme)和10 147个独立的组合述位(complex rheme),如表2统计所示.其中,对于部分组合主位或组合述位,内部还标注了核心(nucleus)和辅助(satellite)两种类型的子主位和子述位,这体现对组合主位或组合述位在语义上的影响重要程度;也体现对下文篇章主位推进过程中可能存在的参与作用.

Table 2 Distribution of Theme/Rheme

表2 主位/述位类型分布

Item	Complex	Nucleus	Satellite	Total
Theme	10 147	1 471	1 929	13 760
Rheme	10 147	324	598	11 069

实体和事件在传统篇章语义分析中占有重要的地位,是组成话题的主要成分.为了体现主位/述位与实体事件之间的联系性,我们在主位/述位标注中从实体事件角度引入了标注属性Type,用来表示某个标注单元属于实体或事件的特点.表3统计了本语料库中,属于实体或事件类型的组合主位和组合述位单元的数量及比例.从中可以看出,属于实体的主位占比远大于属于事件的主位,达到了92.07%.这主要是因为,主位结构的划分,是在小句中的谓词前方内容,体现的是句子的主语成分,采用名词或名词短语这一类实体表示主语的概率相当高.而述位结构的划分,则是包含了谓词及其后方内容,即一般含动宾结构的成分居多,显然属于事件表示的概率要高.

Table 3 Distribution of Event and Entities in Theme & Rheme

表3 组合主位实体事件类型分布

Item	Entity		Event		Total
	Number	Number/Total/%	Number	Number/Total/%	
Complex Theme	9 538	92.07	609	7.93	10 147
Complex Rheme	852	3.57	9 295	96.43	10 147
Total	10 390	51.20	9 904	48.80	20 294

汉语重意合,在子句中会大量出现缺省主语(或宾语等)的情况,因此也带来了包含主语的主位(或包含宾语的述位)的缺省.我们把这种情况称为隐式主位(zero theme)或隐式述位(zero rheme)现象;对应则称为显式主位(explicit theme)或显式述位(explicit rheme).表4统计了本语料库中隐式组合主位和显式组合主位所占比例.

可以看出,隐式主位比例接近30%,占有较大比例;而显式主位比例占比约70%.显式占比大约为隐式占比的2.5倍.此外,表4还统计了隐式组合述位的缺省情况,相对于显式组合述位,占比更低.这主要是因为述位的成分在句法结构而言主要是谓

Table 4 Distribution of Zero or Explicit Theme & Rheme

表 4 组合主位隐式显式类型分布

Item	Zero		Explicit		Total
	Number	Number/Total/%	Number	Number/Total/%	
Complex Theme	3041	29.99	7106	70.01	10147
Complex Rheme	509	5.11	9628	94.89	10147
Total	3550	17.54	16734	82.46	20294

词结构,从词法成分来看主要是动宾结构,而谓词结构组成了一个句子的核心成分,一般而言不会缺省.

3.3.1.3 篇章微观话题结构(MTS)统计与分析

篇章微观话题结构是一个四元组,其构建过程主要包括篇章基本话题中的主位和述位的识别、以及前后基本话题之间的联接识别.在一个篇章微观话题结构中,根据微观话题联接(MTLink)所连接的上下篇章基本话题中的不同主位或述位,可以构成不同的微观话题主位推进模式.

主位推进模式反映篇章的演变规律,是作者表达意图的重要方式.同时,不同体裁的篇章,也通过不同的主位推进模式反映表达风格.如中文小说、散文等,在行文时讲究"抑扬顿挫,曲径通幽",一般不会直接引出主题;而新闻类文章,则为了表达清晰简明的需要,一般会直奔主题,避免"绕弯子".上述不同文体的篇章,在主位推进模式中都有不同体现.

根据定义 10~13,我们在语料库中标注了放射型、集中型、延续型和交叉型共 4 种主位推进模式.

从表 5 所示统计可以看出,放射型主位推进模式占比相当高,达到了 95.75%;其后是交叉型,占比较小,仅约 3.83%;余下 2 种模式所占数量则几乎可以忽略不计.这部分体现了新闻类篇章在上下文衔接过程中的特点,即结构比较简单,主要由主位(主语)结构引导,线性展开文章的意图.

Table 5 Distribution of Thematic Progression Patterns

表 5 主位推进模式分布

Thematic Progression Patterns	MTS-CosTP	MTS-CenTP	MTS-SimTP	MTS-CrsTP	Total
Number	4848	6	15	194	5063
Number/Total/%	95.75	0.12	0.30	3.83	100

3.3.1.4 篇章话题结构(DTS)统计与分析

由第 2 节关于篇章话题结构的定义可知,通过微观话题联接(MTLink)递归连接微观话题结构(MTS),即可以形成篇章话题结构(DTS).微观话题联接(MTLink)体现了篇章的话题演变过程,而主位推进理论中的主位推进模式直观地反映了篇章话题演变关系,将其应用于汉语篇章话题链的识别即可构建一个完整的篇章话题结构体系.

篇章话题链的节点数量及其排列拓扑关系是篇章话题链的主要特征,决定了不同篇章话题链的形态模式,间接反映了篇章的衔接性.表 6 统计分析了本语料库中的篇章话题链节点数量.

链节点的数量多少能够直观地反映作者表述一个话题的复杂程度.很显然,复杂的话题需要加以描述或解释的语句相对要多,进而形成微观话题结构的数量要多,形成多个链节点的概率相对就要高得多.从表 6 可以看出,本语料库中占据多数的节点数量不多,以 2~4 个链节点比例最高,总数约占到 75%,表明本语料库新闻语料所表述的话题总体来看并不复杂,这与新闻语料篇章追求简单快捷、通俗易懂地传播新闻消息的要求是吻合的.此外,从表 6 的统计数据还可以看出,链节点的数量与使用的频率基本上呈现线性递减关系:链包含的节点数量越多,在篇章中使用的频率越少.这体现出本语料中围绕同一个子话题讨论分析的过程不会太久,这也符合新闻语料短小精悍的特点.

Table 6 Distribution of Discourse Topic Chains with Different Number of Nodes

表 6 篇章话题链节点数量分布

Number of Nodes Each Chain	Chain Number	Chain Number/Total/%	Number of Nodes Each Chain	Chain Number	Chain Number/Total/%
2	680	40.05	16	3	0.18
3	402	23.67	17	4	0.24
4	204	12.01	18	7	0.41
5	109	6.42	20	3	0.18
6	91	5.36	21	3	0.18
7	59	3.47	22	2	0.12
8	36	2.12	23	1	0.06
9	24	1.41	24	1	0.06
10	16	0.94	25	2	0.12
11	15	0.88	26	1	0.06
12	15	0.88	27	1	0.06
13	5	0.29	29	1	0.06
14	5	0.29	38	1	0.06
15	7	0.41	Total	1698	100.00

3.3.2 标注一致性分析

在语料标注过程中,尽管不同标注者遵循同一

标注规范,但依然存在由于个体主观性差异而导致标注语料结果的不一致.一致性检验即用来验证这种差异程度,并反映问题的本质难易程度.常用的一致性检验方法是 Kappa 检验.

Kappa 检验借助观察一致率(observed agreement)和偶然一致率(agreement by chance)两个参数来计算用来反映标注语料一致性的 Kappa 值,

$$Kappa=\frac{P_0-P_c}{1-P_c},$$

其中,P_0 表示观察一致率,Pc 表示偶然一致率.Kappa 值$\in[-1,1]$.在评估一致性时,如果 $Kappa>0.75$,一般认为标注具有较好一致性;如果 $Kappa\leqslant 0.4$,则表明一致性较差.为符合常规要求,我们采用 Kappa 方法来检验语料标注质量.

我们以篇章基本话题(子句)为单位,当微观话题结构中的链式结构,即链式结构两端的主位或述位完全相同时,认为微观话题结构的标注结果一致.在语料上分别计算主要标注对象,如篇章基本话题(EDTU)、主/述位(theme/rheme)以及微观话题结构(MTS)的 Kappa 值.表 7 给出了语料库中主要标注对象的标注一致性检验,均值 $Kappa>0.75$,因此,我们认为该语料的标注结果是可靠的;一致性检验表明 CDTC 能够充分体现汉语篇章话题分析问题本身的难度,并能够为相关研究提供语料资源支持.

Table 7 Label Consistency Checking

表 7 标注一致性检验

Recognition Item	Kappa
EDTU	0.91
Theme/Rheme	0.83
MTS	0.81

3.4 CDTC 语料库在自然语言处理领域的应用分析

3.4.1 同类语料库比对分析

从基本单元、联接词、关系表示结构等方面,将我们提出的基于篇章微观话题结构构建的汉语篇章语料库体系 CDTC 与 PDTB 中文标注体系以及汉语广义话题结构体系(GTS)[2]进行比较,结果表明 CDTC 体系吸收了 PDTB 体系和广义话题结构体系的优势,具有合适的篇章话题结构分析粒度,可以满足篇章话题结构分析的需求.具体结果如表 8 所示:

Table 8 Comparison of Chinese Discourse Topic Structures

表 8 同类汉语篇章话题结构体系比较

Item	GTS	PDTB	CDTC
EDU	Punctuation Clauses	Predicate-argument view	An independent punctuation sentence with a predicate verb structure, usually a clause
Topic	Generalized topic and topic clause	PropBank	Micro-topic scheme based on Theme-Rheme Theory
Topic Link	A stack model of the dynamic generation	With each word sense connected to an ontology, and coreference	Micro-topic link based on Thematic Progression
Discourse Structure	Linear superposition	Partial tree, deduced by connective and it's argument	Micro-topic chain; top-down segmentation

语料库的研究,我们认为一般可以分为 3 个阶段:1)利用语料库分析发现语言现象,总结语言规律的过程;2)在此基础上,扩大语料规模,在不同领域验证语言规律的过程;3)进一步扩大语料规模,为具体应用提供充分的语料资源.从研究阶段来看,本文所讨论的语料库资源建设及其语言模型计算,尚处于第 1 阶段.这个一方面是遵循语料库研究的基本规律,另一方面也由于篇章话题结构的复杂性及研究难度,难以快速逾越,还需要持续深入一个时期的研究.

同时,对比参考目前实际面向应用的典型语料库建设来看,在语料规模和覆盖领域 2 个方面都有不同建设特点.例如语料标注规模并非很大的知名语料库就有修辞结构篇章树库、篇章图库等.修辞结构篇章树库 RST-DT 共包含 385 篇文章,由美国南加州大学标注,于 2002 年经 Linguistic Data Consortium(LDC)正式发布,为修辞结构理论 RST 研究提供了研究资源.篇章图库(discourse graph bank,DGB)是根据 Wolf & Gibson 提出的图结构表示篇章的方法加以标注的语料库,共标注了 135 篇文章,用作篇章结构分析的语料资源.

相对而言语料规模比较庞大的典型语料库也有,如宾州篇章树库 PDTB,包括了《华尔街日报》的 2 304 篇文章,于 2008 年正式发布,共标注 4 类篇章

关系. OntoNotes 语料库包含广播和脱口秀节目、新闻、网络日志、电话用语等各种体裁的语料;根据来源,语料可以分为来自英语通讯社、中国通讯社、中国广播新闻、英语广播新闻等,累计包含 290 多万个词. 其中英语通讯社以《华尔街日报》为主,中国通讯社以新华社为主,中国广播新闻主要包括中国中央电视台、中央人民广播电台、中国电视系统等,英语广播新闻也是主流的如美国广播公司、CNN、NBC 的公共国际广播电台和美国之音等,因此能够确保语料来源的权威性.

上述不同规模和覆盖领域的语料库资源,事实上都在自然语言处理的不同研究领域、不同阶段发挥着不同程度的影响和作用.

3.4.2 CDTC 语料库基础应用分析

从后续应用来看,基于我们的篇章话题结构分析结果,在自动摘要、文本分类、信息抽取和机器翻译等领域的应用方法都有应用价值. 比如在自动摘要中,通过话题结构的主述位推进,可以反映话题的变化规律,从而推断作者表达的意图及重点内容,为自动摘要研究提供语料资源. 又如在文章体裁分类中,不同体裁的文章所采用的篇章话题结构推进模式是不同的,其中蕴含着某种结构规律,这个可以为体裁分类提供新的特征. 又如在机器翻译领域,统计翻译方法可以考虑词对齐、短语对齐、子句对齐,那是否也可以基于主述位结构的对齐方法呢? 基于主述位结构的对齐反映已知信息和未知信息、原有话题和新话题的变化规律,能够从篇章层面提供更为准确的语义对齐.

4 结束语

本文提出了一种汉语篇章话题结构的形式化表示模型,并基于此模型构建了汉语篇章话题结构语料库(CDTC). 考虑到标注语料的认可度以及开展篇章衔接性和连贯性联合研究的需要,我们选取了 CTB6.0 中的生语料资源进行标注. 为确保标注的规范性和一致性,我们制定了一整套标注规范,并采用合理的标注策略和人机结合的标注方法进行语料的标注工作. 我们对 CDTC 语料库进行了系统的统计分析和一致性检测,结果表明,该语料库能够较好地反映出篇章话题结构的语言现象和特点,其质量能够达到相关研究对语料的要求. 最后,我们还通过比较同类典型语料库的特点,说明了 CDTC 语料库在基本语料单元、语料库结构等方面具有的优势,以及为自然语言处理应用所提供的重要支撑作用.

目前我们的 CDTC 语料主要来自新闻类文本. 考虑到篇章话题结构的复杂性,我们下一步的研究工作重点将扩大语料标注的规模和文本篇章的类型,以便为篇章话题结构提供更为充分的研究资源.

参 考 文 献

[1] Shang Ying, Song Rou, Lu Dawei. General topic structure theory perspective self-sufficient in topic sentences and study [J]. Journal of Chinese Information Processing, 2014, 28(6): 107-113 (in Chinese)
(尚英, 宋柔, 卢达威. 广义话题结构理论视角下话题自足句成句性研究[J]. 中文信息学报, 2014, 28(6): 107-113)

[2] Song Rou. Chinese chapter generalized topic structure model of the water [J]. Studies of the Chinese Language, 2013(6): 483-494 (in Chinese)
(宋柔. 汉语篇章广义话题结构的流水模型[J]. 中国语文, 2013(6): 483-494)

[3] Song Yang, Wang Houfeng. Chinese zero anaphora resolution with Markov logic [J]. Journal of Computer Research and Development, 2015, 52(9): 2114-2122 (in Chinese)
(宋洋, 王厚峰. 基于马尔可夫逻辑的中文零指代消解[J]. 计算机研究与发展, 2015, 52(9): 2114-2122)

[4] Zhang Muyu, Li Yaobing, Qin Bing, et al. Based on the center word matching refers to dissolve [J]. Journal of Chinese Information Processing, 2011, 25(3): 3-8 (in Chinese)
(张牧宇, 黎耀炳, 秦兵, 等. 基于中心语匹配的共指消解[J]. 中文信息学报, 2011, 25(3): 3-8)

[5] Chao Yuanren. A grammar of spoken Chinese [M]. Berkeley, CA: University of California Press, 1968

[6] Cao Fengfu. Clause and sentence structure in Chinese: A functional perspective [R]. Taipei: Student Book Co, 1990

[7] Qu Chengxi. Chinese Discourse Grammar [M]. Translated by Pan Wengua, et al. Beijing: Beijing Language and Culture University Press, 2006 (in Chinese)
(屈承熹. 汉语篇章语法[M]. 潘文国等译. 北京: 北京语言大学出版社, 2006)

[8] Liu Lijin. Comparative Study Between English and Chinese Discourse Structure Mode [M]. Guangzhou: Sun Yat-sen University Press, 2011: 166-178 (in Chinese)
(刘礼进. 英汉篇章结构模式对比研究[M]. 广州: 中山大学出版社, 2011: 166-178)

[9] Wang Jianguo. A Continuation of the Theory of Topic: Based on the Topic Chain of Chinese-English Discourse Research [M]. Shanghai: Shanghai Jiao Tong University Press, 2013 (in Chinese)
(王建国. 论话题的延续: 基于话题链的汉英篇章研究[M]. 上海: 上海交通大学出版社, 2013)

[10] Zhou Qiang, Zhou Xiaocong. Based on the topic of Chinese discourse coherence description system [J]. Journal of Chinese Information Processing, 2014, 28(5): 102-110 (in Chinese)
(周强, 周骁聪. 基于话题链的汉语语篇连贯性描述体系[J]. 中文信息学报, 2014, 28(5): 102-110)

[11] Xu Jiujiu. Chapter in Modern Chinese Linguistics [M]. Beijing: The Commercial Press, 2010 (in Chinese)
(徐赳赳. 现代汉语篇章语言学[M]. 北京: 商务印书馆, 2010)

[12] Jiang Yuru, Song Rou. Based on the theory of generalized topic sentence recognition [J]. Journal of Chinese Information Processing, 2012, 26(5): 114-119 (in Chinese)
(蒋玉茹, 宋柔. 基于广义话题理论的话题句识别[J]. 中文信息学报, 2012, 26(5): 114-119)

[13] Le Ming. Chinese discourse rhetoric structure tagging research [J]. Journal of Chinese Information Processing, 2008, 22(4): 19-24 (in Chinese)
(乐明. 汉语篇章修辞结构的标注研究[J]. 中文信息学报, 2008, 22(4): 19-24)

[14] Xue Nianwen. Annotating discourse connectives in the Chinese Treebank [C] //Proc of the Workshop on Frontiers in Corpus Annotations Ⅱ: Pie in the Sky. Stroudsburg, PA: ACL, 2005: 84-91

[15] Zhou Yuping, Xue Nianwen. PDTB-style discourse annotation of Chinese text [C] //Proc of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2012: 69-77

[16] Huang Henhsen, Chen Hsinhsi. Contingency and comparison relation labeling and structure prediction in Chinese sentences [C] //Proc of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue. Stroudsburg, PA: ACL, 2013: 261-269

[17] Zhang Muyu, Song Yuan, Qin Bing, et al. Chiese discourse relation recognition [J]. Journal of Chinese Information Processing, 2013, 27(6): 51-57 (in Chinese)
(张牧宇, 宋原, 秦兵, 等. 中文篇章级句间语义关系识别[J]. 中文信息学报, 2013, 27(6): 51-57)

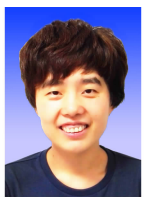
[18] Li Yancui. Research of Chinese discourse structure representation and resource construction [D]. Suzhou: Soochow University, 2015 (in Chinese)
(李艳翠. 汉语篇章结构表示体系及资源构建研究[D]. 苏州: 苏州大学, 2015)

[19] Halliday M A K, Christian M. An Introduction to Functional Grammar [M]. London: Hodder Education Press, 2004

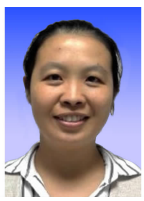
[20] Xing fuyi. The Study of Chinese Sentence [M]. Beijing: The Commercial Press, 2001 (in Chinese)
(邢福义. 汉语复句研究[M]. 北京: 商务印书馆, 2001)

[21] Song Rou. Chinese chapter generalized topic structure model of the water [J]. Studies of the Chinese Language, 2013 (6): 483-494 (in Chinese)
(宋柔. 汉语篇章广义话题结构的流水模型[J]. 中国语文, 2013 (6): 483-494)

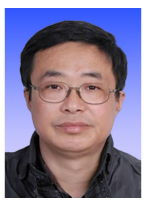
Xi Xuefeng, born in 1978. PhD candidate, associate professor. Member of CCF. His main research interests include natural language understanding, machine learning.



Chu Xiaomin, born in 1981. PhD candidate at Soochow University. Her main research interests include natural language processing and discourse analysis.



Sun Qingying, born in 1982. PhD candidate at Soochow University. Her main research interests include natural language processing, sentiment analysis, stance detection and social computing.



Zhou Guodong, born in 1967. Professor, PhD supervisor. Senior member of CCF. His main research interests include natural language understanding, Chinese computing, and information extraction.