Modern Linguistics 现代语言学, 2017, 5(4), 340-349
Published Online November 2017 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2017.54046

Hans汉斯

Formal Description for the Argumentative Discourse Structure

Maosheng Zhong[1,2], Chao Jiang[2], Qi Wang[3]

[1]School of Computer and Information Engineering, Jiangxi Normal University, Nanchang Jiangxi
[2]School of Information Engineering, East China Jiaotong University, Nanchang Jiangxi
[3]School of Humanity and Social Science, East China Jiaotong University, Nanchang Jiangxi
Email: zhongmaosheng@sina.com

Received: Oct. 24th, 2017; accepted: Nov. 7th, 2017; published: Nov. 15th, 2017

Abstract

With gradually deepened research on the semantic represent and automatically understanding in word or sentence level, the research task on semantic represent and automatically analysis or understanding in discourse level also began to become a focus of research. Because there is close correlation between semantic analysis of discourse and discourse organization structure of text, automatically and accurately obtaining the discourse organization structure of text will be helpful to automatically analyze or understand the semantic of discourse level. In this paper, for the sake of realizing the automatically analyzing for argumentative discourse structure, based on the previous research work and took the Chinese argumentative text or discourse as the object of study, we researched the formal description method for the argumentative discourse structure, put forward some important concepts of the argumentative discourse, such as Element Argument Structure (EAS) and Recursive Argument Structure (RAS), presented a numerical represent method for hierarchical structure, and analyzed and explained how to formalize the argumentative discourse structure by using based on case-analysis method. This study provides a theoretical basis for analyzing the hierarchical structure of argumentative discourse and generating the text structure tree, also has laid the foundation for storing or reconstructing the hierarchical structure of the text.

Keywords

Discourse Structure, Formal Description, Element Argument Structure, Recursive Argument Structure, Text Structure Tree

论证体篇章结构的形式化描述

钟茂生[1,2]，江　超[2]，王　琪[3]

[1]江西师范大学，计算机信息工程学院，江西　南昌

文章引用: 钟茂生, 江超, 王琪. 论证体篇章结构的形式化描述[J]. 现代语言学, 2017, 5(4): 340-349.
DOI: 10.12677/ml.2017.54046

[2]华东交通大学，信息工程学院，江西 南昌
[3]华东交通大学，人文社会科学学院，江西 南昌
Email: zhongmaosheng@sina.com

收稿日期：2017年10月24日；录用日期：2017年11月7日；发布日期：2017年11月15日

摘 要

随着词语、句子等语言单位的语义表示和自动理解研究的逐渐深入，篇章一级的语义表示和自动分析理解也开始成为研究的焦点。由于篇章语义分析与篇章组织结构密切相关，自动准确地获取篇章的组织结构，有助于实现篇章级语义的自动分析。文章为实现论证体篇章组织结构的自动分析，在前人研究工作基础上，以中文论证体篇章为研究对象，研究论证体篇章组织结构的形式化描述方法，提出了论证体篇章中的基本论证结构EAS、递归论证结构RAS等重要概念，同时提出了一种层次结构的数字表示方法，并用实例分析方法来进一步解析如何对论证体篇章进行形式化。该研究工作为论证体篇章的层次结构分析和文本结构树的生成提供了理论依据，也为计算机存储和重构文本的层次结构奠定了基础。

关键词

篇章结构，形式化，基本论证结构，递归论证结构，文本结构树

1. 引言

随着人们对自然语言中词语、句子等语言单位的语义表示和自动理解研究的逐渐深入，篇章一级的语义表示和自动分析理解也开始成为研究人员关注的焦点。由于篇章级语义的分析与篇章组织结构密切相关，如果能够准确地得到篇章的组织结构，将有利于读者或计算机从文章全篇的视角，更好地把握文章的主题或中心思想，从而更准确地分析得到篇章级的语义。

文章从直观上看是一个前后衔接和连贯的句子线性序列。为了实现计算机自动分析篇章的组织结构，首先必须从语言学的篇章结构理论出发，结合计算机进行篇章组织结构自动分析的实现要求，对篇章组织结构进行形式化抽象描述，这是计算机进行篇章结构自动分析的前提，也便于将机器分析的结果进行内部存储和表示。

据此，本文在廖秋忠关于论证体篇章整体结构研究的基础上[1]，以姜岷山的篇章系统性理论为依据[2]，以中文议论文文章为研究对象，研究面向篇章级语义自动分析的篇章组织结构形式化描述方法。下面首先介绍篇章结构研究的相关工作，然后探索提出论证体篇章结构的形式化描述方法，并用实例说明篇章组织结构的形式化描述结果。

2. 相关研究

语言学研究人员对篇章结构理论研究很多，典型的篇章结构理论包括 Mann 的修辞结构理论[3]、Halliday 的语篇性理论[4]、Grosz 的会话结构理论[5]、Kamp 的篇章表示理论[6]、Skorochod’ko 的文本拓

扑结构理论[7]、姜岷山的篇章系统性理论[2]等。此外，廖秋忠分析了论证体篇章的整体结构[1]，列出了汉语论证结构的几种常见类型，它们可以作为识别论证体文章的依据，廖秋忠分析了篇章的局部结构和管界问题[8]，徐赳赳对段落的分布、特性及划分段落的种种制约因素做了一些统计上的考察[9]。不过上述理论或研究基本上是从语言学本身来解释和分析文章的组织结构，虽然对计算机进行篇章结构自动分析具有一定的理论意义，但很难直接用于篇章结构自动分析。

部分计算语言学研究人员，从篇章结构自动分析或篇章结构标注角度进行篇章结构研究。Mann 提出的修辞结构理论(RST)认为一个篇章中的各个小句之间存在着各种各样、不同层次的修辞关系，并对常见的小句之间的修辞关系进行了形式化的描写[3]。乐明根据修辞结构理论，以财经评论文章为语料，定义了标点符号为边界的篇章修辞分析基本单元和 47 种区分核心性单元的汉语修辞关系集，并初步制定了手工进行篇章结构标注的工作守则[10] [11]。刘挺和王开铸给出了一个篇章多级依存结构的形式化描述，篇章多级依存结构不仅展现了文本的逻辑结构，还确定了同层结点之间的依存关系[12]。单永明对文本进行了一个形式化描述，并在此基础上给出了规范文本或准规范文本的篇章结构标引算法，文献中所指的规范文本是指文本结构树中同一层的结点必须同为标题或同为段落[13] [14]。张美娜等是在文献[13] [14]的基础上对文本形式化描述的扩充[15]。

虽然 Mann 和 Thompson 提出的篇章结构树确实能够全面地展现原文的句与句之间、片段与片段之间和章节之间的关系，但是由于这些修辞关系的确定对机器来说本身就是一个困难，而且修辞关系集本身也是开放的，同时，陈莉萍认为，汉语是一种概念耦合型语言，注重意合，用修辞结构理论对汉语进行篇章结构标注时，很难发挥其应有的作用[16]。此外，文献[12]、[13]、[14]和[15]是对一个具有明显的标题、子标题、段落等组成部分的文章进行的形式化描述，其存在的问题是，首先必须识别出标题和子标题以及子标题的级别，这对一些没有明显子标题标记的文章，无法运行其标引算法，而且其算法标引得到的是文本物理结构(非逻辑结构)。

3. 论证体篇章结构的形式化描述

本文在廖秋忠关于论证体篇章整体结构研究的基础上[1]，以姜岷山的篇章系统性理论为依据[2]，进一步对论证体篇章结构进行形式化的描述、修改和扩充，以便为计算机进行篇章结构自动分析奠定基础。下面用正则表达式形式来对篇章文本中的一些概念进行定义，其中的一些符号含义如下：

“$^{=}$”意思是“等价于”或“定义为”；

“+”表示进行变量(分量)的连接运算；

“[]”的意思是“或”，即从方括弧内列出的若干个分量中选择一个，用‘|’号分开供选择的分量；

“{}”的意思是“重复”，即重复花括弧内的分量。如果定义集合，则表示里面是集合中的元素；

“()”的意思是“可选”，即圆括弧里的分量可有可无。

定义 1 汉字字符串集，英文字符串集，数字字符串集，标点符号集，章节符号集

汉字字符串集 CStrings={u| u 是具有实际意义的汉字字符串}；

英文字符串集 EStrings={v| v 是英文单词字符串}；

数字字符串集 Nstrings={w| w 是阿拉伯数字串}；

标点符号集 Punctuations={p| p 是标点符号}。定义中文文本中 P_1={？，！，。…}，英文文本中 P_2={?，!，.，…}，Punctuations= $P_1 \cup P_2$；

章节符号集 TitleStrings={t| t 是章节符号}，例如“一”、“第一章”、“第二章”等。

定义 2 句子(SimpleSentence，简称 SS)，复句(CompoundSentence，简称 CS)

句子$(SS)=u+\{u+([v|w])\}+p$，其中‘+’为连接运算，p∈Punctuations，u∈CStrings，v∈EStrings，

w∈Nstrings；

复句$(\mathrm{CS})=\mathrm{SS}+\mathrm{p}+\{(\mathrm{SS}+\mathrm{p})\}$，其中 SS 为简单句，‘+’为连接运算，p∈Punctuations。

定义 3 标题(Title，简称 T)

□□$(\mathrm{T})=(\mathrm{t})+\mathrm{u}+\{\mathrm{u}+([\mathrm{v}|\mathrm{w}])\}+(\mathrm{p})$，‘+’为连接运算，t∈TitleStrings，p∈Punctuations，u∈CStrings，v∈EStrings，w∈Nstrings；

当标题为文章的题目时，其标题记为 T_0，其他标题记为 T_i，通常 T_0 后面没有标点符号，而其他标题后面可能有标点符号，也可能没有。

定义 4 标题的级(depth，记为 d)

文章中标题的级递归定义如下：

1) 文章的题目(标题)T_0 的级定义为 $\mathrm{d}(\mathrm{T}_0)=0$；

2) 若 T_{ki} 是标题 T_k 的子标题，则 T_{ki} 的级定义为 $\mathrm{d}(\mathrm{T}_{ki})=\mathrm{d}(\mathrm{T}_k)+1$，其中 k_i,k 为标题的下标变量，$0<k_i,k<n$，n 为大于 0 的正整数。

定义 5 自然段(Paragraph，简称 Para)，文章(Text)

$$\text{自然段}(\mathrm{Para})=[\mathrm{SS}|\mathrm{CS}]+\{([\mathrm{SS}|\mathrm{CS}])\};$$

$$\text{文章}(\mathrm{Text})=\mathrm{T}_0+\mathrm{Para}+\{(\mathrm{T}_i)+(\mathrm{Para})\}$$，其中 i 为标题的下标变量；

也可以将文章(Text)看成是一个有序的前后衔接和连贯的句子序列，因此也可以定义为：文章(Text)= [SS|CS] + {([SS|CS])}，在这里，相当于将文章中的所有标题都看成是一个句子，因此文章成为了句子的集合序列。

定义 6 文章主题(Topic)或子主题(subTopic)

文章主题(Topic)指的是文章的中心思想，通常用一个或若干个关键词来表示，有时也用句子来描述文章主题。文章的子主题(subTopic)是相对于上一级主题而言，用来概括当前一个完整语义片段的主要思想或观点，常用一个或若干个关键词或句子来表示。

$$\text{文章主题}(\mathrm{Topic})=[\mathrm{keyword}+\{(\mathrm{keyword})\}|\mathrm{SS}|\mathrm{CS}];$$

$$\text{子主题}(\mathrm{subTopic})=[\mathrm{keyword}+\{(\ \mathrm{keyword})\}|\mathrm{SS}|\mathrm{CS}];$$

其中 keyword ∈ CStrings ∪ EStrings。

下面结合上述定义并参考廖秋忠(1992)，来定义与篇章结构相关的概念。由于本文以中文议论文文章为研究对象，因此这里就论证体篇章结构进行形式化定义，其它体裁文章可相应参考。其中用到的一些符号含义如下：

S 表示文本中的一个结构单元，可以是句子，句群，或段落等单位，

$\mathrm{S}=[[\mathrm{SS}|\mathrm{CS}]+\{([\mathrm{SS}|\mathrm{CS}])\}|\mathrm{Para}+\{(\mathrm{Para})\}]$。S 的下标表示该结构单元在篇章中的序位；

A 表示论证结论，$\mathrm{A}\in\{\mathrm{Topic}\}\cup\{\mathrm{subTopic}\}$；

P 表示论题，包括论点或论断，$\mathrm{P}=[\mathrm{SS}|\mathrm{CS}]$；

E 表示论据，$\mathrm{E}=[[\mathrm{SS}|\mathrm{CS}]+\{([\mathrm{SS}|\mathrm{CS}])\}|\mathrm{Para}+\{(\mathrm{Para})\}]$；

PosE 表示正面论据，$\mathrm{PosE}=[[\mathrm{SS}|\mathrm{CS}]+\{([\mathrm{SS}|\mathrm{CS}])\}|\mathrm{Para}+\{(\mathrm{Para})\}]$；

NegE 表示反面论据，$\mathrm{NegE}=[[\mathrm{SS}|\mathrm{CS}]+\{([\mathrm{SS}|\mathrm{CS}])\}|\mathrm{Para}+\{(\mathrm{Para})\}]$；

I 表示引言，$\mathrm{I}=[[\mathrm{SS}|\mathrm{CS}]+\{([\mathrm{SS}|\mathrm{CS}])\}|\mathrm{Para}+\{(\mathrm{Para})\}]$；

C 表示结尾， $C=\left[\left[SS|CS\right]+\left\{\left(\left[SS|CS\right]\right)\right\}|Para+\left\{\left(Para\right)\right\}\right]$；

CI 表示澄清， $CI=\left[\left[SS|CS\right]+\left\{\left(\left[SS|CS\right]\right)\right\}|Para+\left\{\left(Para\right)\right\}\right]$；

Q 表示问题， $Q=\left[SS|CS\right]+\left\{\left(\left[SS|CS\right]\right)\right\}$；

Ans 表示答案， $Ans=\left[SS|CS\right]+\left\{\left(\left[SS|CS\right]\right)\right\}$；

⌒表示结构单元组件的连接运算；

Δ 表示相关语段的层次结构不再进一步分析。

定义 7 基本论证结构(Elementary Argumentation Structure，简称 EAS)，递归论证结构(Recursive Argumentation Structure，简称 RAS)

基本论证结构是指包括一个论证结论 A、至少一个论题 P、至多一个引言 I、至多一个结尾 C 和若干个论据 E 所构成的论证结构；而递归论证结构则是有若干个基本论证结构递归嵌套而成。两种结构的形式描述如下(用重写规则表示)：

结构 EAS： $A\to\widehat{I\left(P_1\cdots P_m\right)C}$， $P_j\to\widehat{P_jE_1\cdots E_n}$，其中 $E_i\in\left\{PosE\right\}\cup\left\{NegE\right\}$；

结构 RAS： $A\to\widehat{I\left(P_1\cdots P_m\right)C}$， $P_j\to\widehat{P_jE_1\cdots E_n}$，其中 $E_i\in\left\{PosE\right\}\cup\left\{NegE\right\}$；

$I\to A$， $E_i\to A$， $C\to A$；

从上面递归论证结构 RAS 的定义可以看出，一个 I 可以看做是一个基本论证结构并用 A 来重写，一个 E_i 也可以看做是一个基本论证结构并用 A 来重写，直到不能再进一步分析为止，一个 C 同样也可以看作一个基本论证结构并用 A 来重写。上面的定义可以用图 1 的树型结构图表示，图 1(b)中以基本论证结构为单位进行编号，如 A_0,A_1,…等。

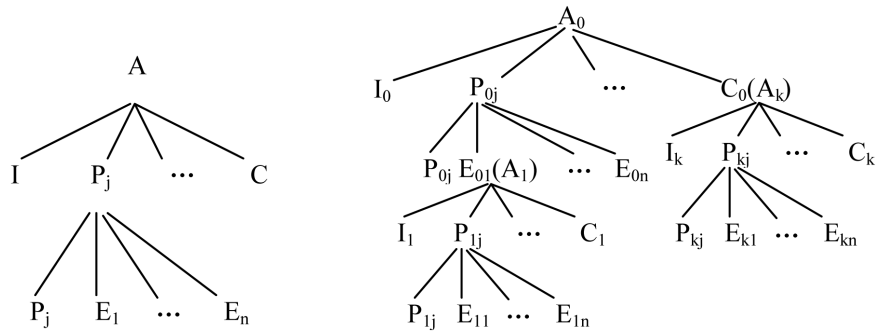
从图 1(a)可知，一个基本论证结构 EAS 通常包括引言、论点和相应的若干论据、结论组成。而图 1(b)中，一个递归论证结构 RAS，为了描述一个引言部分 I，可能又必须先通过一些事例或某种现象接引过来，然后通过论据和论证分析，最后得到引言部分需要的结论，也就是说一个引言部分如果可以再分析的话可以看成是一个基本论证结构。同样，对每个上一层的论据 E_i，又可以看成是下一层的一个 A，因此也是一个基本论证结构。而每个上层的结尾 C，也可以看成是下一层的一个 A，因此同样是一个基本论证结构。通过这种递归形式，可以构造成一个复杂的递归论证结构。这种结构反映了论证体文章可以从不同的视角、不同的层次看到文章的组织结构形式，从最高层可以看到文章只有引言、若干个论点和一个结论部分组成，但是深入下一层后，或许可以对每一部分进行再分析，得到更细化的基本论证结构。

定义 8 文本结构树(Text-Structure-Tree，简称 TST)

文本结构树(TST)是一棵层次结构树，它可以用一个二元组 TST=(*V*, *E*)表示，其中 *V* 为结点集，*E* 为连接结点的边集(树枝集)。令 $V=V_1\cup V_2$，其中 V_1 为 TST 的内结点，是文本对应的递归论证结构 RAS 中的‘A’类型结点的集合，即如果 $v_1\in V_1$，则 $v_1\in\left\{Topic\right\}\cup\left\{subTopic\right\}$；$V_2$ 为 TST 的叶子结点，是文本对应的递归论证结构 RAS 中‘S’类型结点的集合，即如果 $v_2\in V_2$，则 v_2 指向文章中一个句子、一个句群或一个段落。对于 TST 中的任意两个结点 $v_3\in V_1$， $v_4\in V_1$，如果在文本对应的递归论证结构 RAS 树中，v_4 对应的 A 结点是 v_3 对应的 *A* 结点的后代结点，则在 TST 中 v_4 也是 v_3 的后代结点。文本结构树描述的是文本的逻辑结构组织。

例如，如果一个文本有 S_1~S_n 共 *n* 个句子，可以根据上面的定义构造一棵文本结构树如图 2 所示，有 $v_0\sim v_{m-1}$ 共 *m* 个结点，图中空心圆结点表示为内结点，实心圆结点表示为叶子结点，并用箭头线指向实际文本中的句子或句群。

值得注意的是，如果对 TST 从根结点按深度优先搜索进行遍历，并将叶子结点按遍历遇到的先后顺



(a) Elementary Argumentation Structure (b) Recursive Argumentation Structure RAS
(a) EAS 基本论证结构 EAS (b) 递归论证结构

Figure 1. A schematic figure of the elementary argumentation structure and the recursive argumentation structure
图 1. 基本论证结构 EAS 和递归论证结构 RAS 的示意图

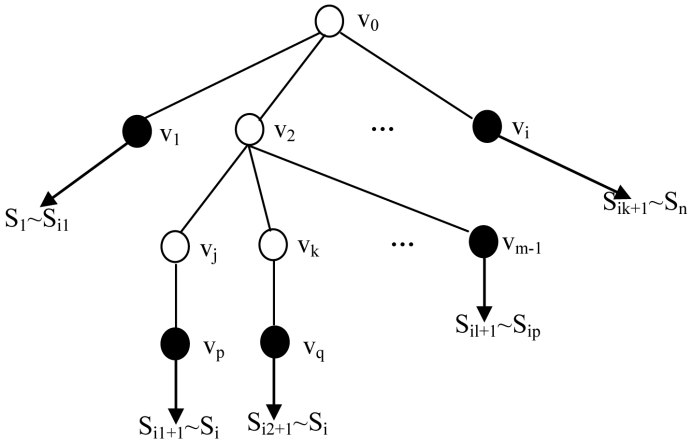


Figure 2. Sketch figure of text structure tree TST
图 2. 文本结构树 TST 示意图

序记录下来，将所有叶子结点所指向的句子或句群按叶子结点的先后顺序进行连接运算，得到的句子序列必须和原文本句子序列相同，因此，叶子结点序列具有偏序性。

文本结构树(TST)表现的层次性可以用括号方法表示，如果对 TST 中每个叶子结点设置两个属性：结点前面的左括号数 n_1、结点后面的右括号数 n_2，则文本层次结构的括号表示方法就可以表示成左右括号数的形式。例如，假设一个文本由 9 个叶子结点(结点可以是段落、句子或句群，这里假定为段落 p_i) $(p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9)$ 组成(其层次结构如图 3 所示)，该文本的组织结构用括号表示法可以表示为：$\left(p_1,\left((p_2, p_3),((p_4, p_5), p_6,(p_7, p_8))\right), p_9\right)$，对每个段落的左右括号数进行统计后，就可以转换成下面的数字表示形式：

$$(1, p_1, 0),(2, p_2, 0),(0, p_3, 1),(2, p_4, 0),(0, p_5, 1),(0, p_6, 0),(1, p_7, 0),(0, p_8, 3),(0, p_9, 1)$$

上面的序列中，每个元素 (n_1, p_i, n_2) 表示段落 p_i 所对应的左右括号数，其中 n_1 表示当前段落 p_i 具有的左括号数，n_2 表示当前段落 p_i 具有的右括号数。因此，文本层次结构的数字表示形式和括号表示形式是一致的，两者之间可以进行相互转换，而且在数字表示形式中，$\sum_{1<i<m}(p_i \to n_1) = \sum_{1<i<m}(p_i \to n_2)$，其中 $p_i \to n_1$ 表示段落 p_i 属性 n_1 上的值。

下面描述 TST 中非叶子结点所对应的主题的级、文本主题单元和主题粒度等概念。

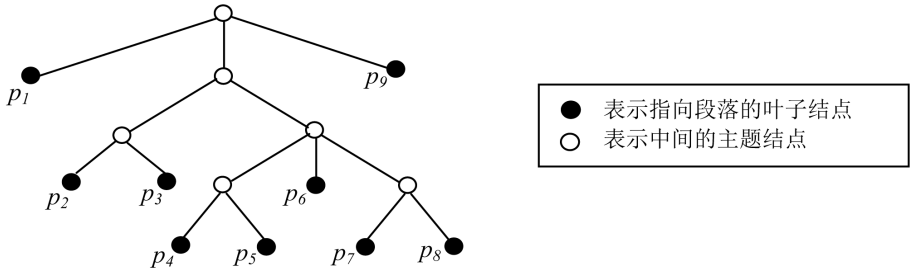


Figure 3. A sketch figure of a text hierarchy structure with 9 leaf nodes
图 3. 一个由 9 个叶子结点组成的文本层次结构示意图

定义 9 主题的级(Level，记为 L)

文本结构树 TST 中非叶子结点所对应的主题的级的定义与文本物理结构中标题的级的定义类似，可用递归的方式定义如下：

1) TST 中根结点 $v_0\left(\in V_1\right)$ 所对应的主题的级定义为 $L\left(v_0\right)=0$；

2) 若 TST 中，v_{ki} 是 v_k 的子结点，并且 $v_{ki}\in V_1$，$v_k\in V_1$，则 v_{ki} 的级定义为 $L\left(v_{ki}\right)=L\left(v_k\right)+1$，其中 k_i,k 为下标变量，$0\le k_i,k\le m$。

定义 10 文本主题单元(TopicUnit，简称 TU)，主题结点粒度(Granularity，记为 G)

文本结构树 TST 中以一个非叶子结点 $v_k\in V_1$ 为根的子树所覆盖的文本块 TU_k 称为一个文本主题单元，记为 $TU_k(v_k\left\langle S_i,\cdots,S_j\right\rangle)$，其中 v_k 为文本块 TU_k 对应的主题。

文本结构树 TST 中文本主题结点(即非叶子结点) v_k 的粒度定义为：$G\left(v_k\right)=2^{-L(v_k)}$。

从上面关于文本主题结点粒度的定义可知，文本结构树中，结点的粒度值 $0<G\left(v_k\right)\le 1$。对于根节点 v_0，由于 $L\left(v_0\right)=0$，所以 $G\left(v_0\right)=1/2^0=1$。此外，主题结点 v_k 所处的级 $L\left(v_k\right)$ 越高，则其粒度值越小，也就是说在 TST 中，根结点 v_0 的粒度最大，其主题概念语义越抽象，而越靠近叶子结点的一些内结点，结点的级越高而粒度值越小，其主题概念语义越明确，这与人对文本中的主题和子主题的概念认知是一致的。还应注意到的是，主题结点 v_k 的粒度是相对于当前文本结构树的一个相对值。

定义 11 文本主题单元 B-I-E 结构

文本主题单元中由一个开始部分(B)、若干个中间部分(I)和一个结尾部分(E)组成的主题内容描述结构称为文本主题单元 B-I-E 结构，其中开始部分(B)是对当前主题内容的一个总的叙述，中间部分(I)则是对当前主题内容分为若干个侧面进行详述，而结尾部分(E)则是对当前主题内容的一个概括性的总结。

对于论证体篇章，文本内一个主题的 B-I-E 结构实际上对应了一个基本论证结构 EAS，如图 4 所示，其中 B-I-E 结构中的开始部分 B 对应了 EAS 中的一个引言 I，而中间部分 I 则对应了 EAS 中的若干个论题 P 和若干证据 E，结尾部分 E 对应了 EAS 中的结尾 C。对于 EAS 中的论证结论 A，可能在 B-I-E 结构中的开始部分 B 出现，也有可能在其结尾部分 E 中出现。文本主题的 B-I-E 结构与基本论证结构 EAS 的区别在于：基本论证结构 EAS 反映了一个文本主题内容内部各部分之间的逻辑或语义关系，而 B-I-E 结构只是从文本组织的外在表现形式上的一种抽象。

从上下文内容相关性的视角看，B-I-E 结构内部的 B 和每个 I_k 均有一定的相关性，E 和每个 I_k 也有一定的相关性。如果将一个主题内容文本看成是由这些结点(B、I_k、E)构成的团，则从直观上看，团内的内容相关性密度较大，而团间的内容相关性密度较小，团的内容相关性密度可用下式进行计算。

$$D\left(\mathrm{B},\mathrm{I},\mathrm{E},k\right)=\frac{1}{2\left(k+1\right)}\left[\sum_{j=1}^{k}\left\{R\left(\mathrm{B},\mathrm{I}_j\right)+R\left(\mathrm{I}_j,\mathrm{E}\right)\right\}+2\times R\left(\mathrm{B},\mathrm{E}\right)\right]$$

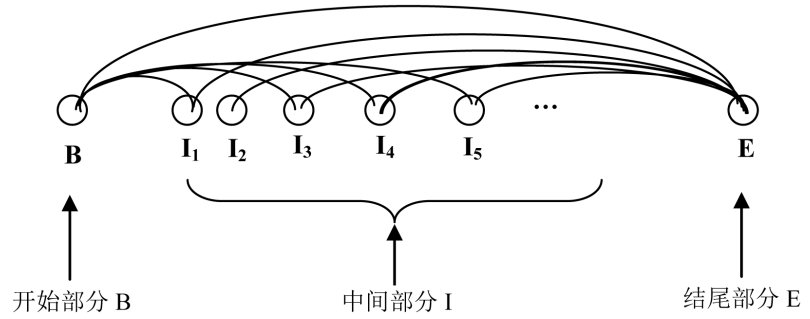


Figure 4. Schematic figure of B-I-E structure for text topics
图 4. 文本主题的 B-I-E 结构示意图

上式中，$D(\mathrm{B},\mathrm{I},\mathrm{E},k)$ 表示由 B，k 个 I 和 E 构成的团 T_i 的内容相关性密度，k 为中间部分 I 的侧面个数，$R(X,Y)$ 表示 X 和 Y 之间的相关度。

4. 示例分析

根据上述篇章结构的形式化描述方法，我们对廖秋忠(1992)中给出的示例文本进行分析，并构建该文本相应的论证结构、文本结构树及其相应的括号表示和数字表示形式、文本主题单元的 B-I-E 结构，该示例文本如下所示。

示例文本：北京乘车难的原因是人多车少——电视辩论赛正方主辨人发言

[我方的中心论点是北京市乘车难的原因在于人与车之间的尖锐矛盾，即客运量大于现在各种公交工具实际的运载能力。[S1]] [这种运载能力既包括车的绝对数量，也包括它的周转速度。[S2]] [兴许大家都有这样的体验，那就是久等车不来，有车上不去，车上人挤人，到站下不来。[S3]] [现在北京人的语言也大大丰富起来了，把“把人挤成了相片”这样形象的比喻可谓是创举了。[S4]] [我方认为，人与车的矛盾从以下这四个方面造成了北京乘车难。[S5]] [第一、北京每天高达 900 万人次的客运量给公交运输造成了难以想象的困难，给社会生活的正常运转带来了沉重的压力。[S6]] [第二、北京现有的运输能力严重不足，长期处于超负荷运转状态。[S7]] [法国巴黎规定公共汽车每平方米最多载客 5 人，伦敦 4 人，而北京呢，早晚高峰时竟达 13 人。[S8]] [第三、北京的运输能力，远远赶不上客运量的加速增长。[S9]] [以承担了 96%客运量的公交公司为例，去年仅有 4033 辆，今年由于报废车辆。绝对数还减少了 72 辆。[S10]] [第四、北京人与车的尖锐矛盾由来已久，积重难返。[S11]] [当然，乘车难还涉及到道路、交通、城市管理、布局等因素，但是这和人与车之间的尖锐矛盾是根本不能相提并论的。[S12]]

该示例文本中有 S_1，…，S_{12} 等 12 个句子，其中 S_1 为论题，S_2~S_4 为对论题的澄清，S_5~S_{12} 为论据部分，包括正面论据和反面论据。由于是一个辩论赛的发言，所以论证结构中没有引言 I 和结尾 C 两部分。而 A 即为“北京乘车难的原因是人多车少”。其论证结构如图 5 所示。

对上述示例文本，如果从宏观上看，由于只有一个论证结构，所以文本结构树上只有一个根结点，该根结点对应的主题即为“北京乘车”，由于只有一个论点，所以对应得到一个“人车矛盾”的子主题，但根据 $\mathrm{E}_i \rightarrow \mathrm{A}$，每个证据可用 A 重写，所以可对图 5 中的 PosE 进行细化和重写，细化和重写后的论证结构如图 6 所示，根据图 6，每个证据 E_i 可对应一个主题，即将原子主题“人车矛盾”进行再细分，可分为“客运量大”、“运力不足”、“运力增长慢”、“人车矛盾久”、“交通资源管理”五个侧面，而“运力不足”、“运力增长慢”属于同一侧面，进行合并。

图 6 中只有一个 A 类型的结点，表明该段文本只叙述了一个主题“北京乘车”，而正方的论点 P 是“北京乘车难的原因是人多车少”(所以可以看做是只有一个子主题，该子主题有四个论据，可将这四个

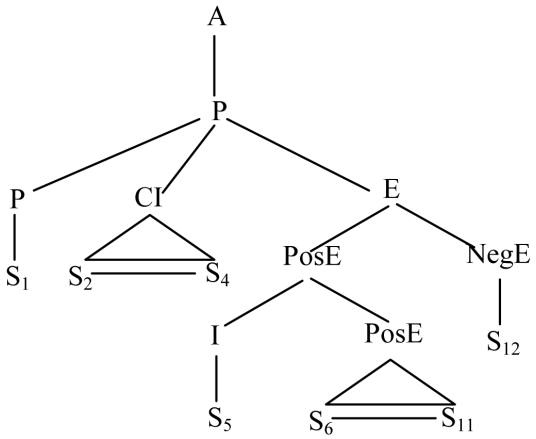


Figure 5. A case of text argumentation structure
图 5. “北京乘车难的原因是人多车少”文本对应的论证结构图

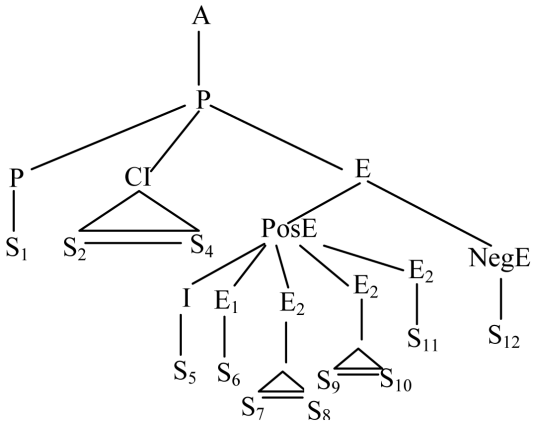


Figure 6. A detail text argumentation structure
图 6. 细化后的论证结构图

论据进行进一步细分为四个侧面，本文不再进行细化)，因此文本结构树中只有一个根结点，此外有一个叶子结点，它指向 S_1~S_{12} 句群，所以可以得到图 7 所示的文本结构树，其相应结构的括号表示为：(p_1，(p_2))，该结构对应的数字表示为<(1，p_1，0)，(1，p_2，2)>。该文本的 B-I-E 结构如图 8 所示，其中 S_1~S_4 为主题的开始部分(为基本论证结构的引言)，而 S_5 为衔接过渡句，S_6~S_{11} 为中间部分，有四个结点，每个结点对应一个论据，S_{12} 为结尾部分(对应基本论证结构的结尾 C)。

5. 结束语

本文为实现论证体篇章组织结构的计算机自动分析，在廖秋忠关于论证体篇章整体结构研究的基础上，以姜岷山的篇章系统性理论为依据，以中文议论文文章为研究对象，研究论证体篇章组织结构的形式化描述方法，提出了论证体篇章中的基本论证结构 EAS 和递归论证结构 RAS 两个重要概念，为论证体篇章的层次结构分析和文本结构树的生成提供了理论依据，同时研究了文本层次结构树中层次结构的括号表示方法，提出了一种层次结构的数字表示方法，为计算机存储和重构文本的层次结构提供了基础。未来的工作主要研究如何自动抽取文本的主题，来实现对论证体篇章的层次结构自动分析，并研究论证体篇章的结构和语义协同分析框架。

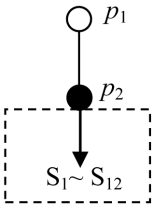


Figure 7. The text structure tree of an example
图 7. 示例文本的文本结构树(图中只有一个根节点和一个叶子节点)

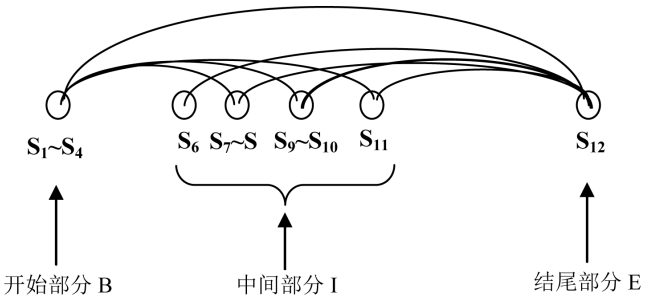


Figure 8. The B-I-E topic structure of an example
图 8. 示例文本的文本主题 B-I-E 结构图

基金项目

本文为国家自然科学基金项目(61462027、61240036)、教育部人文社科基金项目(11YJC740157，09YJC740027)、江西省自然科学基金项目(20114BAB201027)资助研究成果。

参考文献 (References)

[1] 廖秋忠. 篇章中的论证结构[M]//廖秋忠. 廖秋忠文集. 北京: 北京语言学院出版社, 1992.

[2] 姜岷山, 刘汉云, 李学谦. 大学英语篇章结构的基本原理和普遍法则[M]. 北京: 兵器工业出版社, 1993.

[3] Mann, W.C. and Thompson, S.A. (1987) Rhetorical Structure Theory: A Theory of Text Organization. Information Sciences Institute, University of Southern California, Los Angeles.

[4] Halliday, M.A.K. and Hasan, R. (1976) Cohesion in English. Longman, London.

[5] Grosz, B.J. and Sidner, C.L. (1986) Attention, Intentions and the Structure of Discourse. *Computational Linguistics*, **12**, 175-204.

[6] Kamp, H. (1981) A Theory of Truth and Semantic Representation. In: Groenendijk, J., Janssen, T. and Stokhof, M. eds., *Truth, Interpretation and Information*, Foris, Dordrecht, 1-41.

[7] Skorochod, K.E. (1972) Adaptive Method of Automatic Abstracting and Indexing. *Information Processing*, **71**, 1179-1182.

[8] 廖秋忠. 篇章中的管界问题[J]. 中国语文, 1986(4).

[9] 徐赳赳. 篇章中的段落分析[J]. 中国语文, 1996(2).

[10] 乐明. 汉语财经评论的修辞结构标注及篇章研究[D]: [博士学位论文]. 北京: 中国传媒大学, 2006.

[11] 乐明. 汉语篇章修辞结构的标注研究[N]. 中文信息学报, 2008, 22(4): 19-23, 43.

[12] 刘挺, 王开铸. 基于篇章多级依存结构的自动文摘研究[J]. 计算机研究与发展, 1999, 36(4): 479-488.

[13] 单永明. 一类规范文本篇章结构的自动标引[N]. 中文信息学报, 1997, 12(4): 47-51.

[14] 单永明. 汉语文本的篇章结构及其标引算法的研究[N]. 中文信息学报, 2002, 16(2): 14-19.

[15] 张美娜, 迟呈英, 战学刚, 亓超. 基于篇章结构的文本自动标引算法[J]. 计算机应用与软件, 2008, 25(9): 122-124.

[16] 陈莉萍. 汉语篇章结构标注的理论支撑[J]. 南京航空航天大学学报(社会科学版), 2008, 10(3): 68-71.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2330-1708，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：ml@hanspub.org