文本篇章结构的自动标引

张美娜[1]，亓超[1]，迟呈英[1]，战学刚[1]

（1. 鞍山科技大学，鞍山市 114044）

摘 要： 通过对输入文本分析，划分文本的篇章物理结构，分为依次存在包含关系的章节、段落、复句、分句四个层次，用文本结构树来表示。给出了标记方法，并在此基础上讨论并实现了文本篇章物理结构的自动标引，给出了标引算法，为自动文摘后续工作给予了很大帮助。

关键词： 篇章物理结构；文本结构树；标引算法

Automatic Indexing of Discourse Structure for text

Zhang Meina[1], Qi Chao[1],Chi Chengying[1],Zhan Xuegang[1]

(1.AnShan University of Science and Technology, AnShan 114044)

Abstract: In this paper, by analyzing the input text, partition the discourse physical structure of the text, the text is presented as a text structure tree which has four levels of chapter, paragraph, complex sentence and clause that are a inclusive relation. Discussing and implementing the automatic indexing of discourse physical structure for text based on the tagging method and present a indexing algorithm, that is a great help for the further work of the automatic abstraction.

Keywords: the discourse physical structure; text structure tree; indexing algorithm

1 引言

文本篇章物理结构表示了文本的组成情况，在基于篇章结构的自动文摘研究中，为了适应处理大规模真实语料的需要，自动文摘应立足面向非受限领域，而篇章结构属于语言学范畴，不涉及领域知识，因而基于篇章结构的自动文摘方法不受领域限制，同时篇章结构比语言表层结构深入了一大步，但是在进行文本分析之前，如何给出文本篇章物理结构的正确表示是整个工作的关键，为此本文提出了篇章物理结构的两个标引算法，给出了分为依次存在包含关系的章节、段落、复句、分句四个层次，并用文本结构树来表示。

2 基本概念

根据文本的物理结构，一篇文本可以通过文本结构树的形式来表示，可将文本看作是由文本题目、章节、段落、复句、分句按包含关系组成的一个层次结构，而要识别这种包含关系，就需对其给出相应的标记，形成树形结构，致使文本的各个层次可按统一的方式进行访问，根据相应结点在文本结构树中的位置赋予唯一的对应坐标

作者简介：张美娜（1981－），女，江苏射阳人，硕士在读 E-mail:meina1126@yahoo.com.cn

值。

定义 1 汉字字符串集，标点符号集，章节符号集

（1） 汉字字符串集 Strings={u | u 是具有实际意义的汉、英字符串}。

（2） 标点符号集 Punctuations={p | p 是标点符号}。P={？，！，。}，英文文本中 P={?,!,.}，P⊂Punctuations.

（3） 章节符号集 Title={t | t 是章节符号}。例如“一”、“二”、“第一章”、“第二章”，等等。

定义 2 题目和章节标题

设 u∈Strings，t∈Title∪(ε), p 是空串，

（1） 文本题目 h=u。

（2） 章节标题 t，u 的字连接 t·u 称为章节标题 h=tu。

（3） 对于每一篇文本，其中的章节标题应属于同一类型，每一个子集称作 Title 的一个型，这些子集的全体构成 Title 集合。

定义 3 自然段

文本中的一个自然段可递归定义如下：

（1） 若 s=up(up 表示 u 与 p 的字连接 u·p)是句子，且 p∈P，则 s 是自然段。

（2） 若 s=up 是句子（p∈Punctuations），p_i 是自然段，则 s 与 p_i 的字连接 $sp_i=s\cdot p_i$ 是自然段。

定义 4 复句

这里定义复句是由两个或两个以上意思上有联系的分句构成的句子。文本中的一个复句可递归定义如下：

若 s=up(up 表示 u 与 p 的字连接 u·p)是分句，且 p∈Punctuations-P，则 cs=sup（p∈P），则 cs 是复句。

一篇文本由若干个章节组成，每个章节由若干个自然段组成，每个自然段又由若干个句子组成（这里的句子可以是分句也可以是复句），每个复句又由若干个分句组成。

定义 5 文本的标记值

将一篇文本按照其自然结构划分为依次存在包含关系的章节、段落、复句、分句四个层次，用文本的物理结构树表示，根据相应结点在文本结构树中的位置，每一个基本单元赋予唯一对应坐标值，称作文本的标记值，如[1.1.2.1]表示第一章节中第一个自然段的第二个复句中第一个分句。

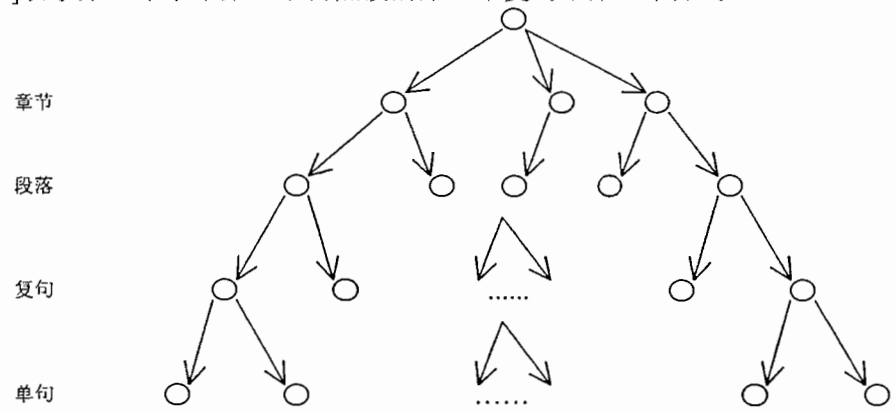

图 1 文本篇章结构示意图

Fig.1 Discourse Structure for Text

3 篇章物理结构标引

根据以上讨论，我们设计了两个篇章物理结构自动标引算法。

textfile：文本文件。存放待分析和标引的文本。文本均以 txt 格式存放，每个基本单元存放在一行中。indexflie：标引文件。用于存放对 textfile 中的文本进行篇章结构标引后形成的标引结果。

算法 1 描述了把文本划分为章节、段落、句子三个层次，这里不考虑复句和单句成分。

步骤 1 初始化 i=0;open (textfile); creat (indexfile)

步骤 2 判断全文标题 (1) read（x，textfile)；(2)若 indent(x)=h 则 B[i].H:=0

步骤 3 每循环一次标引文本其余基本单元

I 若 indent(x)=p，read（next x)，直到 ident (x)=h 为止，当 eof (textfile)=false 反复执行：

II 则(1)i:=i+1;(2)B[i].H: =B[i].H+1； B[i+1].p:=0 B[i+2].s: =0;(3)章节 h，标记为 h B[0].H. B[1].H...B[i].H;（4）append (h , indexfile)；III 若 ident (x)=p 则（1）B[i+1].p: =B[i+1].p+1;（2）段落 p，标记为 p B[1].H...B[i].H. B[i+1].p;（3）append(p , indexfile)；（4）当 eof (textfile)=false 反复执行（每循环一次，标引一个句子)；(i）若扫描到一个 p∈P，则为一个句子；(ii) B[i+2].s:=B[i+2].s+1；(iii)句子 s，标记为 s B[1].H...B[i].H.B[i+1].p. B[i+2].s;(iii) append(s , indexfile)； (5) B[i+2].s: =o

步骤 4 结束

算法1给出了文本篇章物理结构的一个标引算法，但是算法没有划分出复句和分句，所以对所标引的文章的结构太粗，不够精细，看不出文章的内部结构。所以一下讨论算法2给出了复句，分句层次的划分。因为段落中复句之间和章节中段落之间由于存在形式分割标记，容易切分，而复句中各分句之间由于缺乏形式分割标记致使切分难度较高，但是对于形合复句来说，复句中各个分句都是靠关联词来连接的，对于学术类文章，大多数分句间是含有关联词的，所以我们首先将文本分词，然后通过调用关联词词典进行分析，得到了分句的边界标记，同时通过本实验室的语法分析器和指带消解方法拆卸各语言单元之间因省略、指代、引用等造成的内容上的依赖关系，删去关联词语，使每个语言单元在意义上完整独立特别是使每个分句成为自身可理解的命题，这样做是为了最终保证句子的连贯性。

算法2如下：

算法2 描述了把文本划分为章节、段落、复句、分句四个层次。

步骤1,步骤2. 同算法1 步骤1，2

步骤3. 每循环一次标引文本其余基本单元

I, II 同算法1 步骤3 I, II; III （1）（2）（3） 同算法1 步骤3 III （1）（2）（3）;

（4）ii=j=head(段落 p)， eof (textfile)=false 反复执行（每循环一次，标引一个复句和其内的所有分句)：(i) ii=j,j=临近的 p 且(p∈P);(ii)若 ii,j 之间单句标记 c ≥2 则为一个复句,若不是转 II;(iii)B[i+2].cs : =B[i+2].cs+1;(iiii) 复句 cs，标记为 cs B[1].H...B[i].H.B[i+1].p. B[i+2].cs

1≤B[i+3].s≤c ，s[sp+3]++，则分句 s，标记为 s B[1].H...B[i].H. B[i+1].p. B[i+2].cs. B[i+3].s; (iiiii) append (cs, indexfile); append (s, indexfile)；（5）B[i+2].cs: =0;B[i+3].s: =0

步骤 4 结束

在算法中使用的过程或函数说明如下：

open(textfile)：打开文本文件 textfile，文件指针置于文本开始位置。creat(indexfile)：建立标引文件 indexfile。read(x，textfile)：从 textfile 文件指针当前位置开始读取一行。ident(x)：判定函数。若 x 为自然段，返回值为 p；若 x 为标题，返回值为 h。append(x，indexfile)：将得到的标引项 x 内容追加到标引文件 indexfile 中。

4 实验及其讨论

上述算法对于文本结构清晰的文章能够进行有效的划分，我们分别选取了汉语和英语学术类文章各 100 篇作为实验测试集，实验结果如下，

表 1 实验结果

Tab.1 Result of experiment

	100 篇英文		100 篇中文		准确率
	正确篇数	错误篇数	正确篇数	错误篇数	
算法 1	99	1	99	1	99%

算法 2	—	—	80	20	80%

但是对于算法 2 还存在一定的不足，原因是汉语言的复句和分句定义比较复杂，复句的分类主要分为两种：形合复句和意合复句，它们构成的基础都是分句之间的逻辑语义关系。所选的学术类文章中大部分复句中含有关联词，但仍然有意合复句的存在，对于意合复句的划分不够准确，所以要准确划分复句，分句还要从语义层次进行理解，然而汉语语义分析工作还不够完善，所以在这方面还有待进一步的研究。

参考文献：

[1] 王永成. 中文信息处理技术及其基础[M]，上海：上海交通大学出版社，1991.32-50

[2] 王建波[，王开铸]. 自然语言篇章理解及基于理解的自动文摘研究[J]. 中文信息学报，1992， 6(2)：1-7

[3] 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

[4] 刘伟权[，王明会，钟义信]. 建立现代汉语依存关系的层次体系[J]. 中文信息学报，1996， 10 (2)： 32-46

[5] 单永明. 汉语文本形式结构分析及其标引算法[J]. 中文信息学报，2002， 16(2)：14-26

[6] 张益民[，陆汝占，沈李斌]. 一种混合型的汉语篇章结构自动分析方法[J]. 软件学报，2000， 11（11）：1527-1533

[7] G Salton. Automatic Text Structuring and Summarization[J]. Information Processing & Management, 1997， 33(2)：193-207

[8] Salton G[, Allen J. Buckley C]. Automatic Structuring and Retrieval of Large Text Files[J]. Communications of the ACM , 1993 , 37 (2) ： 97 - 108