一种混合型的汉语篇章结构自动分析方法*

张益民, 陆汝占, 沈李斌

(上海交通大学 计算机科学与工程系,上海 200030)

E-mail: lu-rz@cs.sju.edu.cn; yimin.zhang@intel.com

http://www.sjtu.edu.cn

摘要: 提出一种混合型的汉语篇章结构自动分析方法.此方法主要基于RST(rhetorical structure theory)分析、主位模式分析等多种语言学方法,还利用了向量空间模型等统计方法.提出并实现了一种确定性RST分析算法.与其他现有方法相比,此方法具有更大的适用范围和更高的处理精度.

关键词: 汉语篇章结构自动分析;确定性RST分析算法;修辞结构理论;向量空间模型;主位模式分析

中图法分类号: TP391 **文献标识码**: A

汉语篇章结构自动分析是篇章处理中的重要环节,也是大规模真实文本处理系统必不可少的一个组成部分.目前,汉语篇章结构分析的研究主要借助于向量空间模型之类的统计模型[1].这类方法的好处在于适用面较广,对各种类型的篇章均有一定的处理能力.然而这种统计模型有时也显得过于粗糙,对篇章中不少特征信息没有加以很好的利用,因而仅仅基于这种方法很难对篇章中种种细微的结构特征进行很好的处理.

本文提出一种篇章结构自动分析方法,该方法以RST(rhetorical structure theory)理论为其主要的语言学依据,并在其中结合运用了语言学篇章结构研究方面的成果.该方法除了利用连接成分作为求解篇章结构的形式特征之外,还利用了主位模式、汉语篇章管界等更多的有用特征信息来进一步提高该方法的适用范围和处理精度.另外,该方法也利用了启发式规则和向量空间模型方法等来提高鲁棒性.该方法是一种典型的混合方法,因而具有较高的处理精度以及良好的可扩展性和鲁棒性.下面,我们首先给出RST分析的一些理论背景,然后对我们的篇章结构自动分析方法进行介绍,最后给出对该方法的初步的性能测试数据.

1 篇章结构的表示——RS树

1.1 修辞结构理论

语篇(discourse,text)通常是指一系列连续的语段或句子构成的语言整体单位.语篇具有一定的层次结构.语篇中的各个层次的基本组成单位包括小句、句子、段落、章节.语段(text span)是由两个或更多的基本单位构成的、具有结构关系的、功能完整的片段.语段小到由一个句子中的若干小句组成,大到由若干自然段落组成,整个篇章就可以看成是一个最大的语段.篇章就是由以上各种组成单位利用多种关系组合而成的一个相对完整的语言整体.只有分析出语篇中的这种层次结构及其各组成成分之间的语义关系,才能对语篇有一个总体上的把握.

Mann和Thompson提出了修辞结构理论(rhetorical structure theory,简称RST).这是一套关于自然语篇结构描写的理论,重点研究语篇结构的一个主要方面——修辞结构,故名修辞结构理论[2].该理论认为语篇是一个层次结构.在这个层次结构中,无论是小句还是更大的语篇单位之间都是由一些为数不多的、反复出现的关系

* 收稿日期: 1999-05-17; 修改日期: 1999-09-17

基金项目: 国家自然科学基金资助项目(69573020)

作者简介: 张益民(1972-),男,湖南长沙人,博士,主要研究领域为汉语语义理论,自然语言理解;陆汝占(1940-),男,江苏苏州人,教授,博士生导师,主要研究领域为语言模型,自动推理技术,汉语语义理论,自然语言理解;沈李斌(1975-),男,上海人,硕士生,主要研究领域为自然语言处理.

连接.这些关系有时有连接成分作为形式标记,有时则完全是隐含的.RST 理论对小句直至段落间的大量关系进行了详尽的描写和分类.目前,RST 已广泛地用于自然语言生成系统.但是,如何将 RST 用于篇章结构的自动分析尚未得到很深入的研究,而 RST 在汉语篇章结构分析方面的应用研究更是一个全新的领域.

1.2 RS 树的形式化定义

以 RST 分析为基础的篇章结构表示形式一般称为 RS 树,因为它描述的是各个组成部分之间的修辞结构关系,并且是一棵树形的结构.修辞结构关系一般也称为 RS 关系,它表示的是各个组成单位之间的逻辑语义或语用联系,如并列、转折、背景、目的、重言等.下面给出 RS 树的形式化定义.

RS 树是一个六元组$(N,A,RR,NA,\text{Level},\text{BUnits})$,其中

(1) N 是结点的有限集;

(2) A 是弧的有限集,所有的弧都是无向的;

(3) RR 是 RS 关系的有限集,包含了语段间所有可能存在的 RS 关系;

(4) NA 是一个函数:$A\to N\times N$,它定义了弧与所连接的结点对的关系.对 NA 要定义适当的约束规则,使得所定义的结构成为一棵多叉树结构;

(5) Level 表示 RS 树的层次.共有句、段、章节、篇章 4 个层次;

(6) BUnits(basic units)是一个由当前层次上的基本单位组成的序列,记为 $u_1,u_2,\ldots,u_n$.与上面 4 个层次对应的基本单位分别是小句、句、自然段、章节.BUnits 实际上是一个 RS 树的各个叶子结点按中序遍历后所得到的一个序列.

RR 中的 RS 关系分类主要参考了 RST 研究中的关系分类、汉语复句研究中的关系分类以及在篇章连接成分研究中提到的分类.为了使篇章结构分析得以利用形式标记来进行,我们也对 RS 关系进行了综合取舍.我们还为每种关系找到了一些形式标记,大多数形式标记是连接成分,还有一些是管领词语.

2 篇章结构的自动分析

我们给出一个确定性的 RST 分析算法来对汉语篇章结构进行自动分析,其中利用了语料分析的结果和多种语言学特征.

2.1 语料分析

由于我们的 RST 分析将主要基于篇章中出现的连接成分来预测可能的修辞结构关系,连接成分用法的语言学研究中给出了一些基本信息,但这些信息往往是定性的,仍不足以预测可能的篇章结构.因而有必要通过语料分析来进一步获得各个连接成分用法的定量信息和统计信息,其中主要是以下几个方面的信息:

(1) 连接成分的句内外用法识别规则.这些规则主要用于判别一个连接成分在具体上下文中是用于句内小句的连接还是用于更大单位(句子、语段、自然段)的连接.规则获取是通过机器学习方法进行的,类似于 Marcu 所采用的方法[3],但采用的特征集更为广泛.

(2) 用于预测 RS 关系的类型和范围的特征信息.主要有:一个连接成分所表示的各类 RS 关系及其相应的使用频度;关系中各个子语段的大致范围及其对应的使用频度;关系中各个子语段的相对位置及所处的地位.这些信息是通过连接成分用法统计得到的.

(3) 各个 RS 关系的包孕能力.包孕能力指的是各类 RS 关系间的层次优先关系,可以用来解决 RST 分析中的歧义问题.这些信息可在语料分析中通过统计获得.

2.2 RST 分析算法

RST 分析算法是用于篇章结构自动分析的算法.下面,首先对算法的总流程进行说明,然后对算法的各个步骤进行较为详细的介绍.

2.2.1 RST 分析算法流程

RST 分析算法的描述如下:

Input: 文本 T 的句法分析结果.其中包含了每个小句的句法分析树,并经指同求解的处理得到了一些指代

和缺省信息.另外,还包含了小句的边界标记和段落的边界标记.

Output:文本 T 的 RS 树.

算法过程:

Step 1. 根据小句标记、段落标记和标点符号建立篇章的粗结构树,即章节、自然段、句、小句这 4 个基本层面的层次结构.

Step 2. 对章节、自然段、句这 3 个层次分别进行以下步骤:

(1) 识别出篇章中所有与当前层次有关的连接成分和管界成分.

(2) 调用确定性 RST 分析核心算法,求得当前层次的结构树.

Step 3. 利用主位模式分析对以上得到的段、章节层次的结构树进行后处理.

Step 4. 将各层次的结构树合并为一棵完整的篇章结构树.

下面分别对算法中的几个重要步骤进行详细论述.

2.2.2 篇章粗结构树的生成

该步骤主要是利用章节标记、段落标记和一些标点符号来建立一棵粗略的篇章结构树.一个语篇存在着一些自然形成的单位,按单位由小到大的次序为小句、句、段落、章节.这些单位由小到大不断组合,从而形成整个语篇的结构.如果不考虑处于同一层次的各个单位间的关系,这样形成的层次结构可以看成是一个语篇的基本篇章结构.该层次结构非常粗糙,其中句内小句间的关系、段内句间的关系、章节内段落间的关系都没有表示出来,因而我们称之为粗篇章结构.一个语篇的篇章结构可以通过在这个粗篇章结构上进行进一步的分析而得到.

2.2.3 连接成分和管领词语的识别

连接成分主要分为连词、副词、固定短语,只有将它们从文本中识别出来才能进行 RST 分析.更为重要的是,要根据上下文获取连接成分的各种用法信息,以便减少 RST 分析中的歧义,如该连接成分是句内用法还是句外用法、所连接的语段的大致范围等方面的信息.在语料分析阶段,我们已经得到了各个连接成分的用法信息,在这里,可以调用相应的规则来确定该连接成分在具体上下文中的用法.

另外,为了解决篇章管界问题,还要进行管领词语的识别.在篇章中,某些词语,如动词、各种修饰语等具有一定的支配、修饰或统领的范围,当这个范围跨越句子边界时,就称为篇章管界.例如,

例 1:在这些方针指导下,[轻工业的发展加快了,...,这两年来的粮食产量是建国以来最高的,...]

在这个例子中,状语"在这些方针指导下"担任管领词语,它的管界是括在方括号中的部分,由若干个句子组成.篇章管界相对于它的管领词语来说是一种从属的结构关系,因此,管领词语的识别有助于确定局部的篇章结构.具体的识别是根据语料分析中得到的管领词语识别规则来进行的.

2.2.4 RST 分析核心算法

RST 分析核心算法是一个各层次(句、段、章节)共用的算法,只要将某一层次的文本基本单位组成的序列及与当前层次相关的连接成分和篇章管界成分作为输入提供给该算法,就可以构造出该序列所对应的 RS 树.下面,给出它的算法描述.

Input:一个当前层次的基本单位结点组成的序列:$U_1,U_2,\ldots,U_n$,每个基本单位 U_i 中与当前层次相关的连接成分数组 CWD_i 和篇章管界成分数组 DM_i.

Output:该序列对应的 RS 结构树,以 $U_1,U_2,\ldots,U_n$ 为其所有的叶子结点.

算法过程:RSTAna

```
root:=null;
for i=1 to n do
  AddToRSTree(root,Ui);
endfor;
```

其中,变量 *root* 是该序列所对应的 RS 树的根结点,初始时设为 *NULL*.算法的步骤十分简单,主要就是调用 AddToRSTree 将 $U_1 \sim U_n$ 依次加入到 RS 树中去.这是一个渐增式的构造 RS 树的过程,而且也是一个确定性的过程,不进行回溯,每加入一个结点都充分利用各种特征来确定加入树中的位置及如何对树进行重组.因此,该算

法是一种确定性的 RST 分析算法.

函数 AddToRSTree 的作用是将当前单位结点 U_i 加入到目前止在构造的 RS 树中去,下面给出其算法描述.

Input: 当前已生成的 RS 树,根结点为 *root*. 当前正待加入该 RS 树的结点 *tu* 及其中识别出的属于当前层次的连接成分 *cwd* 和篇章管界成分 *dm*.

Output: 加入结点 *tu* 后的 RS 树.

算法过程:AddToRSTree

```
1  mrp[]:=GetMostRightPath(root);{获取最右路径}
2  mrpnum:=size(mrp[]);{获取最右路径上的结点个数}
3  if  sizeof(cwd[])=0{无连接成分}
4     if  dm!=null{有篇章管界成分}
5        havefoundsamedm:=FALSE;{设置同类管领词语出现标志}
6        for  i=mrpnum-1 to 0 do
7           if  deg(mrp[i])>5
8              if  findSameDm(mrp[i].childs[],dm)
9                 havefoundsamedm:=TRUE;
10                reconstructRSTbysameDM(mrp[i],dm);
11       endfor;
12       if  (! havefoundsamedm)
13          nd:=findbestweakmatch(mrp[],tu);
14          AddChild(nd,tu);
15    else  {无篇章管界成分}
16       if  (ReRSTbyParallelStruc()){利用平行句式重组 RST}
17          return;
18       if  (ContinueOrEndGovernSpan()){利用管界结束标志来识别管界范围}
19          return;
20       nd:=findbestweakmatch(mrp[],tu);
21       AddChild(nd,tu);
22 else  {有连接成分}
23    if  (cwd.linkdirection=after)
24       if  (ReRSTbyParallelStruc()){利用平行句式重组 RS 树}
25          return;
26       nd:=findbestweakmatch(mrp[],tu);
27       AddChild(nd,tu);
28    else  if  (cwd.linkdir=before or cwd.linkdir=both)
29       nd:=findfirststrongmatch(mrp[],tu);{找寻第 1 个配对结点}
30       if  (nd!=null)
31          AddChild(nd,tu);
32       else
33          nd:=findbestweakmatch(mrp[],tu);
34          AddChild(nd,tu);
```

以上算法中第 1~2 行获取树的最右路径(从树的最右叶子结点到树根所经过的路径,以该路径上所有结点组成的数组来表示),以便在插入树的时候在最右路径上进行搜索.这种只处理最右路径上结点的做法体现了 RS 树的右斜优先策略[3],因为将结点加入到最右路径上使得所得的 RS 树向右倾斜.第 3~21 行处理当前结点无当前层次连接成分的情形.第 5~14 行处理篇章管界现象,其中第 6~11 行利用 findSameDm 来找到同类管领词语,从而确定结点的插入位置.第 13~14 行则进一步利用弱匹配函数(findbestweakmatch)来找到结点在树中的最佳插入位置.第 16~21 行处理无标记的结点,利用平行结构、管界结束标志和弱匹配函数来找到最佳插入

位置.第23~27行类似地处理连接方向为after的连接成分.这里,连接成分的方向是指连接成分所连接的语段相对于它所在语段的方向.例如,"但是"一般是连接所在语段前面的语段,则其连接方向为before.第29~34行处理需进行连接成分匹配的连接成分,在找不到匹配时,也利用弱匹配来确定插入位置.下面,再对其中使用的几个主要函数作一些说明.

reconstructRSTbysameDm函数的作用是利用同类管领词语再现这种形式特征来确定篇章管界.ContinueOrEndGovernSpan函数主要利用除同类管领词语再现之外的其他特征来确定是继续扩大前面某个管领词语的管界,还是终止前面某个管领词语的管界.

ReRSTbyParallelStruc函数是看是否有平行结构出现,如果有,则要根据平行结构的特点对篇章中存在的层次结构进行重组.

findfirststrongmatch用来在路径mrp[]上找到能与cwd[0]配对使用的连接成分,大多数情况下利用这种信息可唯一地确定当前结点插入树中的位置,因而可称之为强匹配(strong match).

findbestweakmatch则是在没有可适用的强匹配的情况下,利用语料分析所得到的连接成分用法信息及一些歧义消解策略来选取一个较好的匹配结点,与前面的强匹配相对应,我们称这种匹配的方法为弱匹配.具体采用的歧义消解策略主要有:

(1) 指同优先规则.在构造一个篇章的RS树时,将一个新结点加入该树中有多种可能的加入位置,在其他条件均等的情况下,可优先选取一个最近的且其构成短语中(不包含经缺省求解补出的成分)有与当前句中名词短语指同的结点作为加入位置.

(2) 包孕能力优先规则.在构造一个篇章的RS树时,将一个新结点加入该树中有多种可能的加入位置,在其他条件均等的情况下,可优先选取一个不违反包孕能力的加入位置,这就是所谓的包孕能力优先规则[32].具体到各个RS关系间的包孕能力在语料分析阶段已经获得,并存入数据库中,这里,可通过到数据库中查找以获得这些信息.

(3) 向量相似度优先规则.该规则主要是在其他更为确定性的方法不适用的情况下(如段落间无连接成分,也没有列举分承或管界等结构特征),利用向量空间模型的方法进行段间关系的求解.主要思想是把段落表示成向量形式,向量中的分量是该段落中出现的较为重要的关键词语,然后对各个段落根据其向量表示来进行向量相似度比较,采用的比较公式一般是余弦公式.对于一个段落而言,如果它与前面相邻段落的相似度大于某一阈值,则认为是与前面相邻段落成并列关系或阐述关系,并且主题相同,否则可认为它是与前面一连串成并列关系的段落形成并列关系,并且主题不同,这样就可对原来扁平的一层篇章结构分出更细的层次来.具体的阈值设置可以利用文献[4]中介绍的文本分段(text tiling)方法.另外,对于当段落内缺少其他形式特征时的情况,也可以用类似于文本分段的方法,只不过这里的基本单位不是段落,而是一个或多个句子,但算法上没有什么太大的差别.具体细节不再详述.

以上这些优先规则之间也存在优先级的差别,因而对它们的调用有一定的次序,依次是指同优先规则、包孕优先规则、向量相似度优先规则.

2.2.5 RS树的后处理

在一般的语篇中,连接成分等形式标记出现的密度并不小,因而可以用来进行准确度较高的RST分析.但是,也有一些语篇的段落间或段落中的某个局部范围内很少有连接成分出现,这就给RST分析带来了困难,再加上缺少管领词语、平行结构等其他可利用的特征,就更难以判别各个篇章成分间的结构关系.这时,我们将各篇章成分间的关系设置为缺省值:无标记并列关系.它与有标记并列关系不同,它不是真正意义上的并列关系,而只是在信息不足的情况下所做的一种缺省的假设.在已生成的RS树中,由于缺少形式标记,可能会存在不少无标记并列关系.我们将利用主位模式分析来对已生成的RS树进行后处理,主要针对其中的无标记并列关系,从而使RS树的层次结构得以进一步细化.

对RS树进行后处理的算法过程如下.对RS树进行遍历,如果发现某个结点为无标记并列关系,且其度大于某一阈值,则对它的子结点序列进行主位模式分析,并生成相应的主位模式树,然后通过一定的转化规则将主位模式树转化为相应的RS树,从而使该结点的子树结构更为细致.这种更为细致的RS树可更为准确地反映出

篇章的层次结构.

2.3 与其他方法的比较

目前,国外已有一些研究者提出了利用RST来进行篇章结构自动分析的方法,如Marcu[3]和Ono[5].下面,将我们的方法与他们的方法作一比较.

(1) 适用范围不同.我们的方法是一种混合方法,除了利用连接成分之外,还有机地结合运用了主述位分析、平行句式、篇章管界分析、空间向量模型等多种手段来确定篇章结构,而Ono和Marcu的方法则完全依赖于连接成分.因此,我们的方法适用范围更广,对于缺少连接成分的篇章,我们的方法一样能通过其他手段进行篇章结构分析.

(2) RST分析算法的搜索策略不同.Marcu和Ono所采用的方法是,利用穷举法找出与某个语段对应的所有可能的RS树,然后再利用一些简单的启发式特征(主要是右斜特性和关系的层次优先级)来对这些RS树进行评价,根据评价结果选出一棵最优的RS树.而我们采用的RST分析算法是确定性的算法,采用最右路径搜索和多种歧义消解策略来防止不必要的回溯,因而算法的效率要高得多,也使算法的求解正解率更高.

(3) RS树的表示形式不同.我们采用的RS树表示形式为多叉树,而Marcu和Ono则大多采用二叉树.多叉树表示对于表示并列关系等具有多个子语段的关系来说更为方便.

(4) 处理的语种不同.我们的方法主要是针对汉语的,Marcu的方法主要针对英语,而Ono的方法则主要是针对日语.当然,由于我们采用的有些特征(如主位模式)是独立于语种的,因此,我们的方法对于其他语种的篇章结构分析也具有很好的借鉴意义.

本文提出的篇章结构分析方法与国内计算语言学界普遍采用的基于向量空间模型的篇章结构分析方法也有所不同[1].在我们的方法中,向量空间模型只是作为一种辅助手段,主要还是基于多种语言学特征,并且采用的是深层次的特征,因而能进行更为精细和准确的篇章结构分析.

3 实验及讨论

我们对以上的RST分析算法进行了具体实现并进行了初步的性能评价.对某一个语篇的RST分析得到的结果RST树进行评价的方法主要是看其RST评分[3]的正确性,这里主要应看其中的较为重要的一些语段的评分与人工分析结果的相关性,具体做法是,将占原语段中20%、30%、40%的最重要的语段(RST评分越高则越重要)与人工结果相应比例的语段进行比较,得到其Spearman相关系数,该相关系数即反映了所采用方法的性能.我们选取了若干篇章片断,对其人工建立RS树.另外,也有一些是直接从汉语篇章研究专著上获取的,都已有相应的篇章结构表示.然后,将这些篇章片断利用我们的RST自动分析算法进行分析,得到相应的RS树.利用RS树评分方法分别得到人工RS树和自动分析出的RS树中各个结点的权重的偏序排列,根据相应的比例计算二者的Spearman相关系数($p<0.01$),在各个比例上得到的平均相关系数的结果如下:0.634(20%),0.656(30%),0.584(40%).

从结果来看,各个比例上的相关系数都较高(大于0.5).因此,该方法对于篇章的结构分析有较高的正确性.当然,还很有必要进行更大规模的实验,从而对该方法的性能进行更为准确的评价.在该方法的研究中,我们也认识到与语言学界合作的重要性,因此在进一步的研究中,必须与语言学界的专家进行更为广泛和紧密的合作.

References:

[1] Wu, Li-de. Large Scale Chinese Text Processing. Shanghai: Press of Fudan University, 1997 (in Chinese).

[2] Wang, Wei. Introduction to rhetorical structure theory (1). Linguistics Abroad, 1994,1(4):8~13 (in Chinese).

[3] Marcu, D. The rhetorical parsing, summarization, and generation of natural language texts [Ph.D. Thesis]. Department of Computer Science, University of Toronto, 1997.

[4] Hearst, M.A. Multi-Paragraph segmentation of expository text, In: Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Las Cruces: New Mexico State University, 1994. ftp://parcftp.xerox.com/pub/hearst/acl94.ps.gz.

[5] Ono, K., Sumita, K., Miike, K. Abstract generation based on rhetorical structure extraction. In: Proceedings of the International Conference on Computational Linguistics (Coling'94). Kyoto: Kyoto University, 1994. 344~348. http://xxx/lanl.gov/ps/ccm-lg/9411023.

附中文参考文献:

[1] 吴立德.大规模中文文本处理.上海:复旦大学出版社,1997.
[2] 王伟."修辞结构理论"评介(I).国外语言学,1994,1(4):8~13.

A Hybrid Method for Automatic Chinese Discourse Structure Analysis

ZHANG Yi-min, LU Ru-zhan, SHEN Li-bin

(*Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai* 200030, *China*)
E mail: lu-rz@cs.sju.edu.cn; yimin.zhang@intel.com
http://www.sjtu.edu.cn

Received May 17,1999; accepted September 17, 1999

Abstract: A hybrid method is presented for automatic Chinese discourse structure analysis in this paper. This method is based on RST (rhetorical structure theory) analysis, thematic progression analysis and other linguistic methods. Vector space model and other statistical methods are also employed to inhance its robustness. A deterministic RST analysis algorithm is proposed and implemented. Compared with other existing methods, the proposed method has better applicability and precision.

Key words: automatic Chinese discourse structure analysis; deterministic RST (rhetorical structure theory) analysis algorithm; rhetorical structure theory; vector space m