突发事件文本的信息结构分析[1]

曾青青，杨尔弘

（北京语言大学应用语言学研究所，北京 100083）

Email: qing8612@sina.com , yerhong@blcu.edu.cn

摘要： 本文利用RST修辞结构理论研究了突发事件文本的结构关系，重点分析了突发事件类文本中五类主要的结构关系。同时对文本的信息构成作了探讨，分析了客观信息、主观信息以及模糊信息等三类信息，对这三类信息的分布比例也做了一定的研究。这些分析有助于文本标注内容及标注方法的研究，为文本的内容计算研究奠定基础。

关键词： 突发事件 修辞结构理论 结构关系 事件信息 主观信息 客观信息 模糊信息

Analysis of Information Structure in Sudden Events Discourse

Zeng Qingqing, Yang Erhong

(Institute of Applied Linguistics, Beijing Language and Culture University, Beijing 100083)

Abstract: This paper attempts to use Rhetorical Structure Theory to study the structure relations in Sudden Events Discourse, and then pays more attention to five foremost Structure Relations. Meanwhile it discusses the composition of text information, and analyzes three types of information– the subjective information, the objective information and the vague information, and more, it gives some analysis to the distribution. The analysis is useful to the research of the contents and methods of the text annotating, and then it offers the basic research for the text's content computing.

Keywords: Sudden Events, RST, Structure Relation, Event Information, subjective information, objective information, vague information..

1、引言

信息抽取（Information Extraction: IE）是把文本里包含的信息进行结构化处理。目前，计算语言学的热点研究问题就是进行信息抽取，其中一个方向就是发现、追踪并且提取事件信息。

突发事件是超越常规的、突然发生的、需要立即处理的事件，诸如地震、泥石流、火灾、交通事故、恐怖袭击、禽流感、海啸、煤矿事故等。此类事件发生的原因、过程、造成的破坏等都是丰富的信息来源，它本身的性质与机理决定了与他相关的新闻报道是事件信息表达比较突出的文本。从众多的新闻报道中提取出特定的感兴趣的信息，对事件信息进行分析并标注，并以结构化的形式表现出来，就形成信息提取的资源。选定突发事件，对突发事件信息的表达方式、文本所反映的信息成分、结构的规律进行探索，于文本标注十分有益。当前结合RST修辞结构理论进行文本的信息结构分析还不多，因此将研究进行下去，得出结构关系与事件信息抽取之间理论性的结论很有必要。而分析语篇信息的构成类型，讨论主客观信息、模糊信息也有利于对我们的信息提取。

2、突发事件文本的语料来源及选择

关于突发事件的新闻报道很多，用Web网络搜索方式搜索到关键字，从网上下载新闻语篇，然后将html网页转换为txt文本格式，过滤干扰信息，规范文本。

[1]基金资助：国家社科基金项目“面向内容计算的文本信息标注研究”（06YY047）。

突发事件一般包括自然灾害、严重的事故灾害、突发的公共卫生事故、突发的社会安全事件。为了达到研究的全面性，对这四种类型等比例选择有代表性的文本进行分析，依次选取“泥石流”、“地震”、“交通事故”、“煤矿事故”、“禽流感”、“中毒”、“恐怖袭击”、“油价上涨” 八类事件。其中每类文本随机选取20篇，共160篇。运用RST修辞结构理论对每一篇文本进行信息结构分析，分析各小句之间的结构关系，将明显的标志信息标注、收集，并和事件信息抽取建立联系。另外就是分析信息类型，统计客观信息及主观信息以及模糊信息的数量。

3、修辞结构关系与事件信息提取

3.1 RST修辞结构理论简介

RST修辞结构理论由Willam C. Mann 和 Sandra A. Thompson 等于19世纪80年代创立。在英语中，RST以小句（不包括主语从句、宾语从句、限定性定语从句）作为语篇结构的基本单位。在RST理论中，文本的语篇结构可以表示为树：叶节点即语篇的最小单位；中间节点对应着文本的相邻两个片段；节点有核心成分和附属成分，核心成分表现其特征，附属成分提供相关的背景信息，如果不分从属关系，则为多核心的关系；两个或多个相邻且不重合的片段间存在着修辞关系，确定这种关系就确定了一个节点。RST的关系分为单核心的关系和多核心的关系。这些单位之间存在着关系。RST从四个方面对关系定义：1，核心成分（N）的限制条件；2，从属成分(S)的限制条件；3，核心成分与从属成分组合的限制条件；4，效果。

3.2 结构关系的判定与事件信息提取

对于突发事件而言，文章在进行信息表达时，对结构关系的选择运用是有偏好性的。多用背景关系、引述关系、详述关系、非意愿性原因关系、非意愿性结果关系、连接关系。这六个主要的关系在判定时要遵循一定的原则，而它们各自所表达出来的事件信息含量也是不一样的。

3.2.1 背景关系

背景关系是这样判断的，R（读者）在读到S句前不会充分理解N，S的内容在时间上早于N的内容发生或存在。N+S组合后，S增加了R理解N中某一元素的能力。W（作者）意图要达到的效果是R对N的理解增加了。

例3.2.1.1 【[1]自今年3月中旬哥伦比亚进入汛期以来，】【[2]近5万人受灾，】【[3]预计直接经济损失达上亿美元。】

分析：

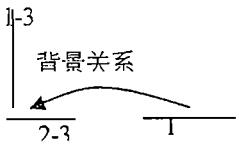

图1 背景关系图解

具有背景关系结构的篇章片段，一般来说从作为从属部分的篇位中能够提取的事件信息是很少的。背景关系中，S的内容在时间上早于N的内容发生或存在，它的出现只仅仅是作为背景知识，增强人们对于N的理解，在提取事件信息时，背景关系的从属成分我们可以不予考虑。背景信息核心成分中一般会包含事件信息，譬如上例篇位2包含了一个受灾事件。

3.2.2 引述关系

引述关系对N没有限制。S是一种言语或认知行为。N+S组合，S揭示了N说话人的身份或信息的来源，且N更为重要。W用这种关系要达到的效果是R意识到S揭示了N说话人的身份或信息的来源且N更为重要。

例3.2.2.1 【[1]巴西民防部门官员1月7日说，】【[2]巴西东南部连日来暴雨成灾，】【[3]引发泥石流，】【[4]已造成约50人死亡。】

分析：

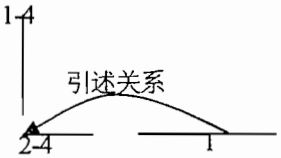

图2 引述关系图解

引述关系，S揭示了N说话人的身份或信息的来源，且N更为重要。在引述关系中，引述的内容才是关注的焦点，间接的客观信息（关于间接客观信息分析见下节）表达很多时候就是用这种引述关系，因而事件信息提取要看核心部分。

3.2.3 详述关系

详述关系对N和S都没有限制，二者组合后，S给出了N中给出的或可以从N中推断到的情景或主题中某一元素的附加细节。这种关系使得R意识到S提供了N的附加细节；R确认主题中的那个元素有附加细节。

例3.2.3.1 【[1]巴基斯坦警方8月29日向媒体表示，】【[2]当天在巴基斯坦南部发生一起重大交通事故，】【[3]一辆超速行驶的大巴士与一辆汽车相撞，】【[4]造成10人死亡，】【[5]26人受伤。】

分析：

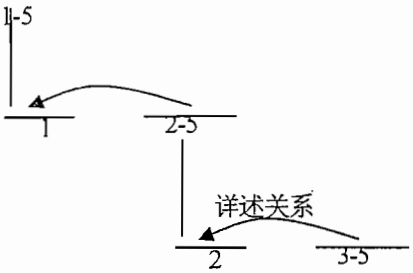

图3 详述关系图解

详述关系的结构段包含事件信息多。不同于背景关系、引述关系，详述关系的核心部分和从属部分都包含事件信息，区别在于核心成分简单精辟，而从属部分提供了大量核心成分的附加细节。这样在面对详述关系时，事件信息提取要从整个结构段入手。新闻报道常常有这样一个特点，正文的第一段一般会对整个事件做一个简单的陈述，而后文在具体描述事件的始末由来。这样的话，在结构树的最上层是一个跨度很大的详述关系，信息抽取在全文展开。

3.2.4 非意愿性原因/结果关系

非意愿性原因关系中，N不是一个意愿性行为，对S没有限制。S并不是通过激励某种意图性的行为而导致了N，没有S的话R可能不知道该情景的具体原因。W意图要达到的效果是让R意识到S是N的一个原因。

例3.2.4.1 【[1]沙尘暴再次袭击呼和浩特市，】【[2]造成呼和浩特机场25个进出港航班取消，】【[3]数千名旅客滞留机场，难以成行。】

分析：

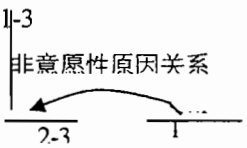

图4　非意愿性原因关系图解

非意愿性结果关系中，对N没有限制，要求S不是一个意愿性行为，并且是N导致了S。这种关系想要达到的效果是R意识到N可能导致了S中的情景。

例3.2.4.2　【[1]位于南太平洋的所罗门群岛海域周一(4月2日)发生强烈地震并引发海啸，】【[2]造成至少18人死亡，】【[3]地方当局已宣布进入紧急状态。】

分析：

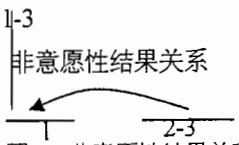

图5　非意愿性结果关系图解

上面两个例子表达的内容很相似，都是A引起了B，但是结合文章标题和内容，用RST修辞结构理论分析发现两句话的强调的核心其实并不一样。

在进行事件信息提取时，这两种关系的核心成分和从属成分都要考虑。无论是强调结果还是原因，都是一个事件引起了另外事件的发生，因此整个结构段都有可以提取的信息。

3.2.5　连接关系

连接关系是一种多核心关系。对N没有限制，对N+N组合的限制是要求几个单元之间的重要性相等，单元间不存在一种已知的修辞关系。而W意图要达到的效果是R意识到在几个单元之间重要性相等但不存在某种已知的修辞关系。

例3.2.5.1　【[1]25日的强震至少造成1人死亡，】【[2]193人受伤，】【[3]建筑物损毁，】【[4]发生塌方，】【[5]并在沿岸引发一次小海啸。】

分析：

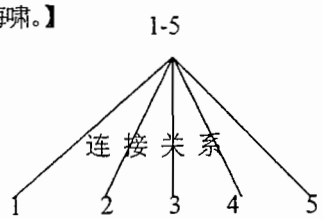

图6　连接关系图解

例中五个篇位都是强震造成的损失，“1人死亡”，“193人受伤”，“建筑物受损”，“发生塌方”，“沿岸引发一场地震”，它们是并列的。连接关系往往由很多篇位并列在一起构成，往往会给出大量的信息。

4、文本信息构成分析

4.1　信息类型

从语言的形式表达和意义阐述两个方面综合来看，事件文本的信息可以分为三种类型，即客观信息、主观信息和模糊信息。客观信息是直接描述新闻事件，或转述事件的客观情况，不带任何主观色彩或评价的信息。因为它只对事件的发生、发展过程做直白的描述，所以其语言形式表达的一个鲜明的特征就是很平实。客观信息经常会告诉我们发生了什么、发生的时间、发生的地点、和事件相关的人物情况、事件的解决途径方法等等。主观信息是文本撰写者或新闻报道的记者对客观事实的来源、成因及发展趋势做的主观评价、判断与推测，或者转述其他人对事件的评价、推测等。其特点在于它是表明人对事情的主观感受或情感介入，表达作者或转述他人的一种个人判断和观点，因而感情色彩很浓厚。语言形式上，多用主观性的词语。语篇分析者面对文本

信息经常会遇到这样一种情况，对于一句话，它是主观信息还是客观信息有一种模棱两可的感觉，无从判断。这种信息介于主观信息和客观信息之间，属于主客观信息间的“中间地带”，由于不好对这类信息划分一个明确的归属，因此将其归为一类，叫做模糊信息。模糊信息因其特殊性，在形式表达方面不好归纳其特征。

在文本当中，有些客观信息是对于事件的一个直接描述，我们称之为直接事件，它属于文本要报道的一个原事件。而很多时候，文本经常会采用一种他人转述的方式来报道事件，是一种间接的事件叙述方式，例如“据消息人士透露”，“记者了解到”，“记者获悉”，“据 xxx 反映”，“xxx 说”，“xxx 介绍”，这些标记性的话语后面所带的信息就不再是原事件，它们与原事件不再同一个层面上，这些结果段和原事件之间是一种信息补充关系，或者有时候是证实关系。对于主观信息也是一样的，它可以是直接的，也可是间接的。

4.1.1 客观信息

在新闻报道突发事件文本当中，很多的客观信息是直接阐述出来的，它们都属于是直接的客观信息。例如这段话：“昨日下午近 3 时，厦门大生里殡仪馆旁山体滑坡，泥石流砸垮一座 2 层的殡仪馆宿舍，一妇女被压废墟中约半小时后被救出，无生命危险。”有的突发事件报道文章全篇都采用一种直接的信息表达方式。

间接的客观信息是通过对其他人的话语转述，或者通过他者将一些客观事实阐述出来。文章在介绍一些事件相关信息时，为了叙述的方便，会采取一种间接的信息表达，例如用“记者了解到”，“据了解”，“记者获悉”，“xxx 表示”，“据 xxx 反映”，“xx 说”等。例如“警方说，一名 52 岁妇女被困坠落的屋顶下，被送往医院后身亡。石川县消防和灾难紧急措施厅说，石川县至少有 154 人受伤，在邻近石川县的富山县，至少 16 人受伤。” 这段话，分别以“警方说”和“石川县消防和灾难紧急措施厅说”这种间接转述的方式将地震造成的人员伤亡事实客观的表达出来，是间接的客观信息。

4.1.2 主观信息

在突发事件文本中，同客观信息一样，主观信息的表达也是有直接和间接之分的。直接的主观信息就是文本的撰写者不通过他人口吻，直接将主观信息用文字表达出来，也即说主观信息的表达不是采用“人+动词”（这些动词表猜测、判断、评价，或者只是一个纯粹不带感情色彩的动词“说”字）的形式。例如“伊朗驻联合国官员的一系列可疑行为已引发了纽约警局官员有关伊朗特工可能主使发动恐怖袭击的担心，”这句话显然是直接的主观信息表达。

很多时候，主观信息的表达采用这样的方式，“记者认为”，“xxx 认为”，“xxx 猜测”，“xxx 说”，这种方式就是间接的主观信息表达。文本常常会转述其他人对事件的评价、推测，以丰富文本的信息，或者增强读者对于事件的了解，例如“另有消息来源指，这是一起非常严重的事件。”

4.1.3 模糊信息

模糊信息是信息类型中的一种特殊状态，例如“当地官员说，地震在当地引起恐慌，1400 多名居民从家中跑了出来”，其中“当地官员说，地震在当地引起恐慌”者可以说是一个客观的事实，但是又可以说是报道者的一种主观判断，分析起来就有较大的难度，只能归属于模糊信息中。

在面对模糊信息的时候，可能不同的语篇分析者会有不同的处理结果，因为判断主客观信息类别很大程度上带有分析者的一种经验性在其中。对于同样一句话，有的人就断定它是客观信息，而有的分析者恰好又认为它是主观信息，还有的人会因为无从判断而将其归入模糊信息。因此，对于模糊信息的界定，还有待于我们进一步研究，给出一个明确的界定方法。

4.2 文本信息分布方式

在讨论文本信息分布方式的时候，模糊信息有其特殊性，故暂不考虑。而在考虑分布方式的时候，不再具体区分直接性和间接性。

4.2.1 纯客观信息文本

纯客观性文本中没有一点主观的内容在里面，这种突发事件文本目的就是将客观事实报道出来，展现给读者看。也就是说，这种新闻报道文本告诉读者一件客观发生的事件，其信息传递通过其所报道的事实的逻辑力量实现。文章采用客观陈述的方法，不渗入一点纯主观的信息评价在其中。例如下面这个例子，就是纯客观信息的罗列：

本报讯　昨日下午近3时，厦门大生里殡仪馆旁山体滑坡，泥石流砸垮一座2层的殡仪馆宿舍，一妇女被压废墟中约半小时后被救出，无生命危险。

该民房依山而建，被泥石流砸垮后，二层的水泥板及家具门窗等压在一楼地板上。停在民房前的一辆大众小轿车车前挡风玻璃被砸破，车身毁坏严重。

在选中的八类事件中，每一类都有这种纯客观性的文本，说明这种客观陈述的报道方式是媒体记者报道采用的一种重要方式。

4.2.2 主客共现型文本

更多的文本采用的是一种主观信息和客观信息共现的方式。对任何一个突发事件，报道者经常会不可避免地加入自己或者他人的一种主观态度，或者说是引用他人的话语来对事件做一个评价、对事态的发展做一个预测，这些都可以说是本来事件的相关信息。大多数新闻文本用的是这种方式，这也是符合行文需要和读者接受习惯的。将主客观“混融”在一起，在叙述客观事实的基础上加上主观信息可以更好更全面地了解整个突发事件。如下例：

越南又有两省出现禽流感疫情，疫情演变复杂。

据报道，越南南部朔庄省美秀县美香乡近几日有66只鸭子先后病死，该省兽医部门14日表示，疫点的全部134只鸭子的样本经H5N1型禽流感病毒检验结果呈阳性。

这个例子中，小句“疫情演变复杂”作为主观信息散布在整篇文本当中，和客观信息一起建构成了完整的语篇。

4.3 主客观信息及模糊信息分布比例

一般来说，统计直接的主客观信息时是以修辞结构理论的小句来计算的，即一个小句算作一个统计计数。对于间接的主客观信息，前面的标记词语不作为统计的成分进入比例分析当中，而其后的句子则同直接的主客观信息一样按照小句结构进行统计。

但在统计时，修辞结构关系有其特殊性，同样的结构关系，有时候组成它的各个小句能表达一个完整的信息，有时候却不能，这样就需要区别对待。例如让步关系，往往是两个或多个小句共同表达一个完整句意。这种情况是以一个完整句子进入统计数量。例如“截至昨晚8时，搜救人员已找到被掩埋压塌的推土机，但仍未找到王某，搜救工作仍在进行中”中，“截至昨晚8时”“搜救人员已找到被掩埋压塌的推土机”和“但仍未找到王某”三个小句必须看成是一个直接的客观信息。而像下面这种情况“记者获悉：两名下落不明的矿工已找到一个，但已死亡”中，“两名下落不明的矿工已找到一个”和“但已死亡”就应该做两个间接的客观信息。同样的，背景关系、非意愿性结果关系/原因关系等也做类似处理。下表是各类文本中各种信息的分布情况。

表1　客观信息、主观信息及模糊信息分布表

信息类别 \ 文本类型		泥石流	地震	交通事故	煤矿事故	禽流感	中毒	恐怖袭击	油价上涨
客观信息	直接信息	110	119	131	94	81	197	79	135
	间接信息	61	77	59	73	74	77	117	84
主观信息	直接信息	8	9	12	3	10	10	7	22
	间接信息	7	4	9	4	13	10	27	29
模糊信息		7	8	0	20	6	9	9	10

从上面的统计数据可以看出，八类文本客观信息的分布明显多于主观信息，这是因为事件报道，人们关注焦点还是事件本身的信息内容。事件发生的时间、地点、伤亡人数，造成的损失、危害，灾害发生后的抢救措施等等都是文本所主要要交待的客观信息；各类文本中，主观信息分布都比较少，而且相对分布比较稳定。

一般直接的客观信息多于间接的客观信息，但在选取的20篇“恐怖袭击”文本中，间接的客观信息明显多于直接的客观信息，这是因为“恐怖袭击”事件涉及国家安全，需要严谨，报道者又害怕承担报道失误的责任，因而不去直接讲述事件，多采取“xxx说”，“另有消息指称”，“xxx透露”，“xxx在一份声明中说”等间接形式。直接的主观信息和间接的主观信息分布比较均匀。

模糊信息所占成分并不多，一般来说，还是很容易对信息的主客观性做界定的。

5、结语

本文一方面用RST修辞结构理论研究了突发事件文本的结构关系，重点分析了突发事件类文本中五类主要结构关系的判定方法及它们与事件信息抽取的联系。另外一方面对文本的信息构成作了探讨，分析了客观信息、主观信息以及模糊信息等三类信息，并对它们的分布比例也做了一定的分析。

以上这些研究，对面向内容计算的文本信息标注研究很有帮助。用RST理论得出关于结构关系的分析后，可以明确信息抽取的侧重范围。而区分了主观信息、客观信息、模糊信息，就可以确定在进行文本信息标注时哪些信息可以先标注，哪些信息可以放入一个待定集中暂不标注。三种信息以及它们的直接性与间接性，对于突发事件信息中区别直接事件与相关事件是十分必要的。随着这些研究的展开，突发事件的信息抽取工作会更为准确、客观，而我们的应对能力也将随之提高。

参考文献

[1] 王伟. “修辞结构理论”评介（上）.《国外语言学》，1994，第4期：p8-p13；

[2] 王伟. “修辞结构理论”评介（下）.《国外语言学》，1995，第2期：p10-p16；

[3] 杨尔弘，邹红建. 面向内容计算的意义单元及其标注研究. 内容计算的研究与应用前沿[A]. 北京：清华大学出版社，2007. p301-p306；

[4] 邹红建，杨尔弘. 以事件标注为核心的语篇标注研究. 第三届全国信息检索与内容安全学术会议报告.

[5] 黄国文.《语篇分析概要》. 长沙：湖南教育出版社，1988. 221；

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

[7] 徐军等. 使用机器学习方法进行新闻的情感自动分类.《中文信息学报》2007，21：p95-p100。