事件标注及突发事件文本内容分析[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@126.com

摘要：本文在事件标注的基础上，对新闻报道中事件报道的组织方式进行分析，界定了文本的主副线信息链，对文本中机器可以识别的内容与难识别的内容进行了分析统计，是信息提取研究的基础工作。

关键词：事件标注　事件属性　主线信息链 副线信息链

Event Annotation and Analysis about the Content in Sudden Events Discourse

Zeng Qingqing　Yang Erhong

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

Email:qing8612@sina.com,yerhong@126.com

Abstract: This paper attempts to analyze the setup about the sudden event discourses. All the analysis is based on a large scale of events annotation. It also defines two information chains from the perspective of Researchers---the main information chain and the secondary information chain. The article also performs a statistical analysis between the content which can be easily identified by the compute and which can not, but Anyway, all the research are just to lay the foundation for Information Extraction.

Keywords: Event annotation, Event attribute, Main information chain, Secondary information chain

1　引言

事件信息抽取是把含有事件信息的非结构化文本以结构化的形式呈现出来，在自动文摘，自动问答，信息检索等领域有着广泛的应用。目前国内外对事件进行信息抽取的研究很多，如MUC、ACE、TimeML、EventML等。综合考察这些研究发现，在识别事件时，抽取范围、信息的散落位置等的确定仍有较大的困难。以ACE为例，它虽然在信息提取方面比较深入，但是它只标注其定义了的几类事件，而放到真实的文本中去考察，只局限于几类事件会丢失很多丰富的信息。

本文从大规模的事件标注入手，论及事件标注过程，进而分析文本内容构成，为信息自动抽取服务。

2　事件标注及事件属性

事件是现实世界所发生的引起特定人关注的事情及其在语言中的表述。它属于组合意义单元，由事件词、事件论元等组成，即通常意义上所说的由动词或动词性概念来描述的行为及包含有时间、地点、参与者等等在内的行为特征。在文本中，事件依据自身的地位又具有特有的属性。

2.1　事件标注

在标注中，我们参照了ACE事件标注的做法[1][2][3]，但是又不局限此。标注是一个扩充集，可

[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）。

以在标注的过程中不断完善修正，以便为事件的信息抽取奠定更为全面的基础。在对事件表达进行标注时，目前采取的策略是按标点句逐句进行标注。标注的主要内容是：事件词、事件词对应的事件类型（子类型）、论元、论元参数以及事件相关属性等。

2.1.1 事件词

事件词是标志该事件发生的词语。例如，“甘肃景泰发生一次Ms4.2级地震”，这里的事件词为“地震”，“夫妻俩是中毒而亡”，这里的事件词为“中毒”和“亡”。对事件表达进行标注时，唯一的依据就是事件词。一般来说，只要在句中出现事件词，就认为该句表达了一个事件，对于一个句子中出现多个不同事件的事件词，则分别标注出这些事件表达。

在真实文本中，一个句子可能会出现多个代表同一个事件类型的事件词，这时候我们可以任意选择一个作为事件词，例如“震中位于关东地区的茨城县海面，震源深约50公里，包括首都东京在内的关东部分地区有一定震感。”此时我们在“震中”、“震源”、“震感”中选择一个。

有的时候，读者明确知道一个句子中发生了某个事件，但却没有出现相关词语让我们可以选为事件词，例如“帝王酒楼厨房一片狼藉，天花板基本上化为灰烬”，这句话明显有一个天花板遭到“毁坏”的事件，但是没有合适的动词，对此类情况，我们标记为事件词缺省。

2.1.2 论元参数和论元

论元是事件标注的一个重要组成部分。在标注系统中，事先对事件类型所做的一个标注模型中会有 arg0、arg1、arg2……argn（n>=0），这些就是论元参数。一个标准的事件有完整的以事件词为原点的论元结构，而事件和语篇具体结合后，文本的构成成分往往是一个典型事件的全部论元参数在其中的部分映射。例如一个标准的“地震”类事件，其关联的论元参数有“arg0 时间”、“arg1 地点”、“arg2 震级”、“arg3 震中”、“arg4 震源”、“arg5 震源深度”、“arg6 有震感地点”、“arg7 持续时间”、“arg8 地震烈度”等，而在具体文本“甘肃景泰发生一次 Ms4.2 级地震”中，文本关联的论元只有地点“甘肃景泰”和震级“Ms4.2 级”。

论元参数和语篇结合后，具体表现形式主要有实体、时间。实体（Entity）是信息提取中的一个基本的概念，对实体的标注我们直接采用ACE（2005a）的实体标注规范[1]，并根据实际标注的需要做了一些调整。实体是指现实世界中的一个对象或者对象的集合。实体表达（Entity Mention）是指指称该实体的语言表达。在文本中，专有名词、普通名词、名词短语以及一些代词都是实体表达。“印度西北部拉贾斯坦邦一座著名清真寺 11 日晚遭炸弹袭击”中的实体有“印度西北部拉贾斯坦邦”、“一座著名清真寺”、“炸弹”。时间表达是为了告诉人们某事何时发生、或者持续多长时间等信息的，包括时点和时段。例如“莫斯科时间 15 日 14 时 50 分许”，“当地时间下午 5 点 40 分”，“当地时间上午 7 时 25 分(北京时间 10 时 25 分)”。

2.1.3 事件属性信息

在标注过程中，我们还会对事件属性信息加以标记，包括事件的模态（Modality）、事件的时态（Tense）、事件的极性(Polarity)、事件的普遍性（Generality）、事件的程度（Extent）、事件的体态（Aspect）、事件的结果（Result）、事件的次数（Frequency）等。

事件模态（Modality），即事件发生的可能性。因为文本在陈述事件时，并不都是肯定的，有些是预测或者估计，事件模态有“确定”和“其他”两类。事件时态（Tense）是指事件发生的时间与文本时间的关系，分为“过去”、“现在”、“未来”、“非特定”四类。事件极性(Polarity)指事件表达的是肯定还是否定，例如，“所幸未发生人员伤亡”，这里的事件的极性就是否定。事件普遍性（Generality）指事件是指向具体某个事件还是某类事件，也就是“一般”和“特指”

两个类别。事件程度（Extent）分为“轻微”、“中等”、“强烈”三大类。有的文本在报道一类事件时会强调该事件的作用程度，例如“印度尼西亚东北部沿海21日傍晚发生强烈地震，印度地震机构测量震级为里氏6.7级”，事件程度为“强烈”。事件体态（Aspect）包括“进行”、“完成”、“非确定”等，例如“目前搜救活动仍在继续中”就蕴含了一个“进行”体态。事件结果（Result）分为“成功”、“未成功”、“非确定”三种情况。有的文本会强调事件的结果信息，例如“消防员将被困人员成功地解救出来”，“解救”结果为“成功”。事件次数（Frequency）是指有的文本会涉及次数，例如“清真寺附近发生了两起爆炸，所幸无人受伤”，“爆炸”事件次数为2次。在标注中，如果文本并没有特意强调事件某一个属性信息，则不标注相关信息。

2.2 事件类别属性

目前，我们人工标注了4类文本，即“地震”、“火灾”、“中毒”、“恐怖袭击”，从中得到了大量的事件，每个文本标注好的事件词构成了这个文本的相关事件词的词集。

显然这些词集里的事件词在文本中的地位和作用是不同的。在一篇新闻报道中，文本的话题决定了事件的重要程度，与文本话题关联最为密切的事件为文本的核心事件，其他为相关事件。

2.2.1 核心事件

核心事件词是标志核心事件发生的词语。所谓核心事件是突发事件文本的重点报道对象，它与文本内容性质相关。譬如，地震类文本的核心事件是“地震”，火灾类文本的核心事件是“火灾”，中毒类文本的核心事件是“中毒”，恐怖袭击类文本的核心事件是“袭击”。在标注过程中，我们很容易发现，核心事件在文本中一般并不是一次提及，往往文本会反复的提及核心事件。

2.2.2 相关事件

相关事件词是标志相关事件发生的词语。突发性事件语篇，除核心事件外，还会有很多其他的事件一起来架构语篇本身，即为相关事件。每个事件文本是不同的，关联的相关事件也会不同。从这个意义上来说，相关事件具有无限可能性。但是在这种无限可能性中我们可以发现某一类文本经常共现的相关事件。例如地震类文本，与“地震”事件共现的事件有“海啸”、“受伤”、“死亡”、“毁坏”、“救援”、“调查”等。又如“中毒”类事件文本，与核心事件“中毒”共现的相关事件有“病态”、“死亡”、“住院”、“治疗”、“调查”等。

3、文本内容分析

从信息抽取目标出发，对于文本的内容，我们需要做一个更为结构化的分析。篇章中并不是所有的信息都适合作为抽取的对象。Van Dijk的话语宏观结构论在抽取最上面的宏观结构层次的过程中有4个基本程序：①去掉不重要或次要信息，②选出基本信息，③把信息加以概括，④选出主题或主题句。[4][5]我们抽取信息时适当借鉴了这一理论，以此出发，我们发现突发事件文本是一个主副线信息结构关系交织网。突发事件文本在信息抽取，提取所需要信息时，主线信息链是我们保留的结果，而副线信息链我们暂时放入暂不标注信息集中，以待以后更方便地研究。

3.1 主副线信息结构

突发事件文本是由主线信息链和副线信息链交叉而成。一个独立的语篇的主线信息链是由以事件词为显性标记的各类事件关联而成。副线信息链则是由一些过细节的描述性信息以及包括背

景信息、主观信息在内的非事件信息等构成。

主副线信息链之间各自独立存在，又紧密地连结在一起，共同形成语篇二位一体的结构框架。

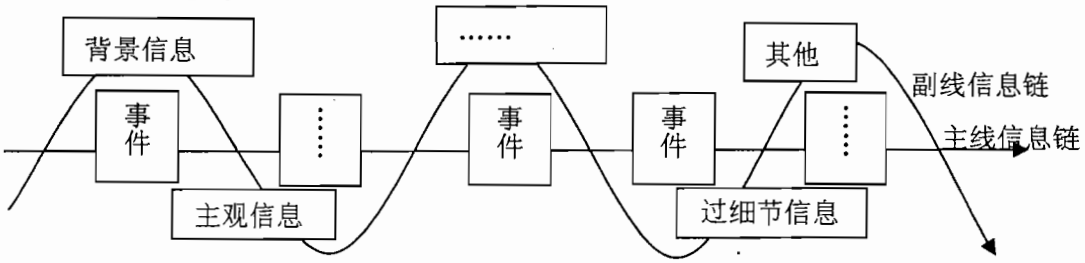

图 1　突发事件文本信息结构链

3.1.1　主线信息链

主线信息链是文本的中心部分，构成了篇章结构的主要部分，是读者进行篇章阅读和理解的最重要部分。它是由上文所述的核心事件和相关事件串联而成的，而这种串联的方式则是通过事件关系。

我们在标注的过程中，也会标注事件关系。事件关系的标注，实质就是确定事件之间的串联关系，从而确立主线信息链。文本之间的事件会通过一些事件类的关系在上下文之间进行很好的整合，从而连贯衔接成为一个整体。在突发事件类文本中，会有很多的事件关系。

3.1.1.1　事件关系

一，因果关系：指事件A的发生导致了事件B的发生。事件文本目的是向人们阐述事情本身的变化情况，所以由“地震”、“火灾”、“中毒”、“恐怖袭击”等事件引发的“受伤”、“死亡”、“伤亡”、“毁坏”、“受困”等事件成为新闻报道的主体内容。二，同指关系：指事件A和事件B指代相同的事件。三，顺序关系：所谓的顺序关系是指在事件A发生之后，在一定长度的时间段内，事件B跟随事件A发生。突发性事件发生并且造成巨大损失过后，依照事态发生顺序，语篇自然会向读者讲述“救援”、“调查”、“部署”、“诊治”等事件。四，解说关系：解说关系是一种比较特殊的关系。新闻文本中经常有一类间接客观信息，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，例如“地震预警中心说，日本今天发生6级地震”，“说”是传递消息类事件，它和“地震”事件构成解说关系。五，并列关系：事件A与事件表达B是同时发生的，或者由同一事件引起的。六，目的关系：事件A的发生是为了事件B的发生。七，转折关系：即事件A虽然发生，但是可能发生的事件B并没有发生。八，其他关系：以上关系除外的关系。

在已标注的各种事件关系中，因果关系最多，其次是同指关系。突发事件往往会造成灾难性后果，因此因果关系在突发事件文本中占据多数。文本还会通过贯穿于文本始终的具有同指关系的事件来表现一种强文本关联。

3.1.1.2　事件关系文本举例分析

中新网 7 月 9 日电综合报道，当地时间 8 日，秘鲁南部发生里氏 6.2 级地震[1]，造成一名 93 岁男子死亡，另有至少 4 人受伤。

据美国地震局H称[1]，这起地震[2]发生在当地时间 8 日上午 4 点 13 分左右。

秘鲁当地官员称[2]，一名 93 岁男子在家中被倒塌的墙壁压死，一家烟花工厂在地震[3]中爆炸，有两名工人被烧伤。

上面这篇文本中，突出显示的为事件词，代表相应的事件。其中“地震 1”、“地震 2”、“地震 3”为同指关系；第一段中“地震 1”和“死亡”、“受伤”为因果关系；“称 1”和“地震 2 ”

构成一种解说关系；“称2”的解说内容包括“倒塌”、“压死”、“地震3”、“爆炸”、“烧伤”5个事件，而5个被解说的事件之间又有自身的事件关系，“倒塌”和“压死”为因果关系，“地震3”和“爆炸”，“爆炸”和“烧伤”为因果关系。

3.1.2 副线信息链

在标注过程中，我们发现，除去主线信息链关联的事件内容，还有很多内容不好标注，但是它们又构成了文本不可缺少的一部分，充实了文本的内容。

3.1.2.1 过细节信息

过细节信息是对事件过于详细的描述，有时候是一种人物语言的引用。一般说来，在标注过程中这类信息因为内容较为琐碎，情节性过强，不易标注。例如下面这个例子：

“今天上午，记者在梁平县仁贤镇中心卫生院四楼的住院部看到，初一（3）班谢同学躺在病床上，双手捂住腹部喊痛，泪水不停的流，院方没有给谢同学打点滴，也没有做任何护理。只有同班另一个女同学正在照顾她，并不停的安慰谢同学。”

很多时候，过细节信息是为了丰富主链的信息成分，使得整个语篇更具生动性和形象性。

3.1.2.2 背景信息

背景信息是辅助语篇受众对象完整和充分理解语篇内容的信息。背景信息分为事件背景、知识背景、历史背景。它或者是对与核心事件具有相似性的事件的概括性提及，或者是对核心事件或相关事件涉及的论元做一些相关的解释或补充描述。

当地时间3日早8时35分，尼加拉瓜西部太平洋海域发生里氏4．8级地震，虽然震中在海上，但尼加拉瓜西部沿海多个城市均有震感。目前此次地震尚无人员伤亡和财产损失报道。地处太平洋板块和加勒比板块交界处的尼加拉瓜境内地壳运动频繁[1]，历史上曾多次发生地震[2]。

以上语篇加下划线的部分即为背景信息，其中小句1是对前文所述核心事件“地震”的地点论元“尼加拉瓜西部太平洋海域”的一个补充描述；而小句2涉及了“地震”事件，这是与本文关注的核心事件具有类同性质。背景信息往往是一种有知识补充作用，帮助人更好地理解文本。

3.1.2.3 主观信息

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。有时候是人物心理活动的一种猜测。

例如“有消息来源指出，这是一起非常严重的事件。”

“虽没有人受伤，但有几名工人想到逃生一幕还是心有余悸。”；“据报道，这起火灾是亨廷顿市近50年来最为严重的一次火灾。”

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。

3.1.2.4 其他

在真实的文本中，还有很多情况不是以上三种能够全部概括的，我们把它归结为其他情况。

例如“罗周忠因外出不在家，逃过一劫，罗还有一个女儿在外地读书。”

“据了解，2116次属于武铁武汉客运段，该列车晚点1小时50分后，2116次列车安全抵达襄樊站。其间，途经石门至长沙间铁路线的部分列车出现了不同程度的晚点。”

“具体报告将在二十九日发布。”

3.2 主副线信息标注情况统计分析

通过人工标注的方法，对4类突发事件文本，每篇选20篇共计80篇文本进行信息可标注性的分析研究，得到的表如下：

表1 突发事件文本主副线信息统计分析表

信息 / 事件类	主线信息		副线信息			
	核心事件	相关事件	过细节信息	背景信息	主观信息	其它
地震	87	142	3	14	3	1
火灾	82	154	15	11	2	5
中毒	74	165	45	7	3	3
恐怖袭击	83	128	10	8	12	4
共计	326	589	73	40	20	13
比例	30.7%	55.5%	6.9%	3.8%	1.9%	1.2%

从此表我们可以看出，突发事件文本中核心事件和相关事件占据了绝大多数文本的内容，副线信息链关联的内容相对而言很少。这个统计也充分证明，以核心事件和相关事件为中心成分的主线信息链才是文本信息抽取的重要关注对象。

4、结语

当今社会，各类突发事件不断，相应的突发事件报道也不断。从这些众多的新闻报道中有效地提取出突发事件的相关信息，需要对文本的语篇结构进行一定的形式化研究，抽取出其中基本的文本构成模式。研究发现突发事件文本中，由核心事件和相关事件在事件关系作用下串联而成的主线链结构涵盖了众多的文本信息，是易于抽取的对象。而一些过细节信息、主观信息、背景信息等副线信息由于其在篇章中的灵活性和多样性而成为了信息抽取的干扰信息。

只有通过对大量真实文本的标注，我们才能不断完善标注体系，才能更加合理地确立文本内抽取范围，过滤干扰信息，从而进一步为文本信息的自动抽取奠定基础。

参考文献

[1] ACE. ACE Chinese Annotation Guidelines for Entities (Version 5.5) [EB/OL].

http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-Entities-Guidelines_v5.5.pdf. 2005a.

[2] ACE. ACE Chinese Annotation Guidelines for Relations (Version 5.5.1) [EB/OL].

http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-Relations-Guidelines_v5.5.1.pdf.2005b.

[3] ACE. ACE Chinese Annotation Guidelines for Events[EB/OL].

http://www.ldc.upenn.edu/Projects/ACE/docs/Chinese-Events-Guidelines_v5.5.1.pdf. 2005c.

[4] 钱敏汝，戴伊克的话语宏观结构论（上），《国外语言学》，1988年第2期，北京；

[5] 钱敏汝，戴伊克的话语宏观结构论（下），《国外语言学》，1988年第3期，北京；

[6] 娄开阳，试论新闻语篇的结构要素系统，信阳师范学院学报第28卷第3期，2008年6月；

[7] 邹红建，杨尔弘，突发事件信息的标注研究，北京语言大学硕士论文，2008年6月；

[8] 钱敏汝，《篇章语用学概论》，外语教学与研究出版社，2001年12月第1版；

[9] 廖秋忠，《廖秋忠文集》，北京语言学院出版社，1992年10月第1版.