基于信息结构的突发事件文本事件信息自动抽取策略研究[1]

曾青青[1]　杨尔弘[1]　朱丹青[2]

（1、北京语言大学应用语言学研究所　北京　100083;

2、中国农业银行 北京 100161）

Email:qing8612@sina.com,yerhong@126.com, zdq0309@gmail.com

摘要：事件信息抽取是自然语言处理中一个较新的领域，汉语方面还有很大的研究空间。本文提出了一个以完整突发事件文本为处理对象的事件信息抽取策略。目前语义角色的识别和分类多是基于句子，针对一个完整语篇进行信息自动抽取更加复杂。我们的方法是通过主副线信息结构链的构成分析，确立了突发事件文本三级层次的事件框架体系，并在此基础上首先过滤副线信息链，然后识别事件词，进而进行论元及事件属性的识别以及分类，形成结构化的数据。

关键词：事件抽取　信息结构　识别　分类

Research of Automatic Event Information Extraction Strategy based on Information Structure Analysis of Sudden Events Discourse

Zeng Qingqing[1]　Yang Erhong[1]　Zhu Dangqing[2]

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

2、Agricultural Bank of China, Beijing, 100161)

Email:qing8612@sina.com,yerhong@126.com, zdq0309@gmail.com

Abstract: Event information extraction is relatively a new field in Natural Language Processing. There is still a lot of space to pay attention to Chinese. This paper presents a event automatic extraction strategy to handle a whole discourse. Via the analysis about the chain of the main information structure and the secondary one, we establish an event frame system which has three levels. On this basis, we firstly filter the vice-line information chain, identify the event words, and then identify and classify the argument and the attribute of the event so as to form a well-structured data.

Keywords: Event annotation, Information Structure, identify, classify

1　引言

在信息膨胀、信息总量迅速增长的时代尤其需要相应的技术来帮助受众耗费较少的时间，以期快速准确地分析信息，定位信息焦点，获取篇章意义的表达。信息检索、信息抽取、自动文摘等NLP技术在这种需求下迅速发展。信息抽取是指从语篇中抽取出预先设定目标的各种事件信息以形成结构化数据。在国外，信息抽取的研究进行地比较早，研究水平比较先进。在国内，信息抽取作为一个较新的领域还有很大的研究空间。

当前和信息抽取相关的语义角色识别和分类工作很多都是基于句子的分析研究，对一个完整的语篇实现事件信息自动抽取要比句子复杂很多。我们发现，突发事件文本的信息结构由主副线信息结构链构成，文本内容主要由主、客观信息构成。本研究在分析事件文本信息结构的基础上，对完整语篇中事件信息抽取整个工作的策略做了一个规划。

2、突发事件文本文本分析

2.1　相关概念解释

词是最小的能够独立运用的语言单位。概念是词在一个特定的上下文环境中所表达的某个确定意思。“义原”是在知网（HowNet）中提出并定义的，是描述各种概念的基础，

[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）。

是经过高度抽象的、不易再作分割的汉语语义最小单位。所有的概念都可以分解成各种各样的义原。知网考察并筛选出有限的义原集合，通过义原的各种组合，描述简单或复杂的概念，进而描述概念与概念、概念与属性、概念的属性与属性之间的各种关系。

2.2 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。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。

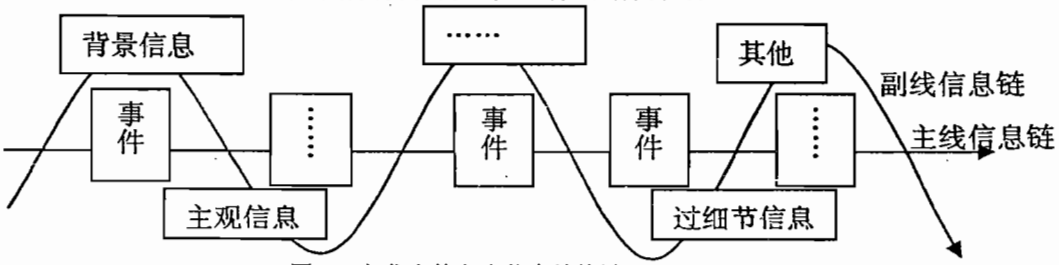


图 1　突发事件文本信息结构链

2.2.1 主线信息链

一个语篇的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。主线信息链由以事件词为显性标记的各类事件通过事件关系串联而成。事件词是标志事件发生的动词或名词。

通过分析我们发现，主线信息链的事件架构是一个三级层次的事件框架体系，包括核心事件本体事件群，核心事件次生事件群以及核心事件再生事件群。后二者构成相关事件。

图 2　突发事件文本三级层次的事件框架体系

以地震类突发事件文本为例，我们通过 python 脚本语言程序从网页上下载并随机选取了 200 篇地震文本，共计 270kb，43,857 字符的语料进行考察。通过对 200 篇文本的人工标注，我们获得了三个事件群包含的事件词。

（1）地震文本中核心事件本体事件群。专指“地震”事件。核心事件本体事件群的事件词是以核心事件为代表的事件词集合：Core Event Words of Earthquake=【地震、强震、大震、主震、微震、弱震、余震、震波、震感、有感、震央、震中、震级、震源、震深、震度、烈度】。这些词绝大部分在 HowNet 中拥有共同的 DEF，即“{mishap|劫难:cause={shiver|颤动:agent={land|陆地}}}”，因而可以用来代表同一个核心事件。

（2）地震文本中核心事件次生事件。指由“地震”造成的诸如“死亡”、“毁坏”等不可抗拒的事件，是直接影响与后果。核心事件次生事件群的事件词是以次生事件为代表的事件词集合：Secondary Event Words of Earthquake=【<摇晃、晃动、摇动、晃、震动>；<伤亡、死亡、遇难、死、失踪、受伤、伤、重伤、轻伤、划伤、跌伤、擦伤、割伤、砸伤、烧伤>；<损坏、受损、破坏、摧毁、毁坏、毁损、损毁、损伤、被毁、破碎、破裂、碎裂、开裂、裂缝、震塌、震碎、震落、跌落、倒塌、坍塌>；<受灾、海啸、洪水、泥石流、滑坡、垮塌、爆炸、火灾、大火、起火>；<损失、被困、惊吓、惊恐、恐慌、避险、离开家园、疏散>；<停电、停水、断电、中断、瘫痪、停运、停驶、关闭、影响>】。

（3）地震文本中核心事件再生事件群。指“地震”发生后，灾难造成的间接影响，主要是人面对灾难采取的各种应对措施。地震文本第三层——核心事件再生事件群的事件词是以再生事件为代表的事件词集合：Regeneration Events Words of Earthquake=【<启动（应急预案）、警告、预警、警报、撤销（预警）、取消、解除>；<调动、调拨、指挥、派出、派遣、赶赴、赴、奔赴、前往、赶往、抗震、救灾、抗震救灾、救援、抢修、支援、搭建、发放、运送、救助、搜寻、搜索、营救、救护、抢救、治疗、动员、转移、疏散>；<调查、分析、监测、检测、检查、排查、核查、监视、勘察>；<承诺、哀悼、慰问、呼吁>】。这些再生事件基本上由“预警”、“救援”、“调查”以及“安抚”等几部分的事件构成。

2.2.2 副线信息链

该信息链条包括过细节信息、背景信息以及主观信息等。过细节信息、背景信息以及主观信息等在内的副线信息是为了丰富主线信息链以外的信息成分，使得整个语篇更具生动性和形象性，帮助读者更好地理解文本，获取全面的信息量。

过细节信息是对事件过于详细的描述，有时候是引用人物语言。其特征在于内容琐碎，情节性过强，语言较为口语。例如：“一家服装店店员说：‘我一开始不知道发生了什么事。我突然感到地面在晃动，货架和墙也都在晃。我赶紧跑到大街上，街上到处都是人。我们现在都还没缓过来。’”背景信息是辅助语篇受众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，或者是对文本事件涉及的论元做一些相关的解释、描述。例如：“冰岛是个岛国，这里2000年6月曾发生里氏6.6级地震，造成一些房屋倒塌，但没有造成伤亡。冰岛气象部门地质学家埃纳尔·基亚尔坦松说，这次地震是2000年以来最强烈的一次地震。冰岛地质活动频繁，间歇泉和火山闻名遐迩。”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。”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。

2.3 信息构成类型

从语言的形式表达和意义阐述两个方面综合来看，事件文本主副线信息链的信息类型可分为三种，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。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。在文本当中，客观信息和主观信息有直接和间接之分。文本常通过直接描述给出信息，很多时候，文本也会采用间接的方式来给出信息。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”“记者认为”“xxx认为”“xxx猜测”“xxx说”等就是间接信息表达。语篇分析者面对文本信息中的某句话经常会无从判断其主客观性，这种信息属于主客观信息间的“中间地带”，叫做模糊信息。

3、突发事件文本信息抽取策略

3.1 问题的形式化描述

(1)给定：一个描述某类突发事件的汉语事件语篇$T=\{S_1, S_2, \cdots, S_n\}$，其中$S_1, S_2, \cdots, S_n$是T中的n个有序句子（以句号、叹号等为标记）。$S_m=\{S_{m1}, S_{m2}, \cdots\}$（m=1, 2…, n），其中$S_{m1}, S_{m2}, \cdots$是每一个句子的小句集合，小句的数量不定但是有限。

T中抽取的事件定义为$E(<R_1, R_2, \cdots, R_i>, <A_1, A_2, \cdots, A_j>)$，其中，$R_1$，$R_2$，…，$R_i$为事件E的i个论元角色，$A_1, A_2, \cdots, A_j$为事件E的j个事件属性。

(2)目标：从T中识别出事件E的各个论元角色R_1，R_2，…，R_i以及文本中提及的各个事件

属性$A_1, A_2, \cdots, A_j$并进行分类。

(3)突发事件文本事件信息抽取包括信息过滤、预处理、识别、分类、后处理等步骤。如下图所示:

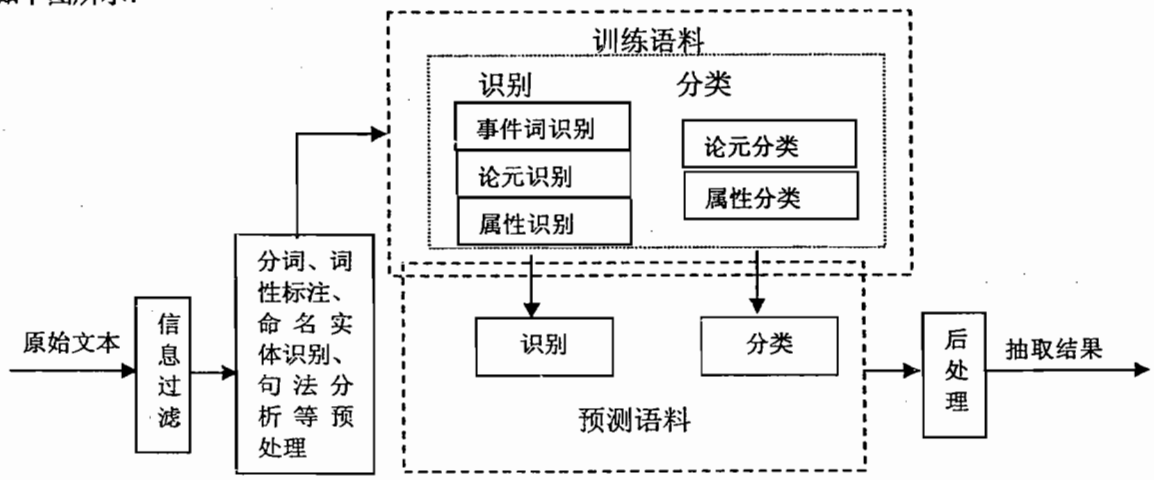


图3 突发事件文本事件信息抽取步骤

3.2 信息抽取步骤具体描述

3.2.1 信息过滤及分词、句法分析等预处理

(1) 信息过滤。之所以过滤是为了尽可能增加下一步识别工作的准确度。所谓信息过滤是指过滤文本中影响事件抽取的干扰信息。我们希望利用特征将副线信息链的信息提取出来放入暂不标记的集合当中。

背景信息是比较好过滤的信息，主要是利用词语显性标记作为过滤手段。我们对200篇地震文本考察，发现很多背景信息表达方式如下："日本地震频发，每年发生有感地震1000多次，是世界上地震最频繁的国家之一。""2007年8月15日，秘鲁中部地区曾发生过7.9级地震，造成595人死亡，300多人失踪，7.5万所住宅被毁。"从中我们找到诸如"(频繁)|(频发)|(多发国家)|(多发区)|(多发带)|(多发地带)|(强地震带)|(最易发生)|(经常发生)|(活跃)|(曾发生)|(曾多次发生)|(曾遭遇)|(发生过)|(上次发生)|(上一次发生)|(去年)"这样的词语显性标记。在我们选取的200篇地震文本中，人工标记有59个句子是背景信息。通过perl程序将这些词语作为抽取显性标记，能够抽出45个句子，抽全率为76.27%。

未抽取出来的句子诸如"呼图壁县位于新疆中北部，距离新疆首府乌鲁木齐约六十公里。""巴达赫尚省是阿富汗最偏远的地区，交通不便、通信落后、人口密度很低。"等，需要找另外的规则特征进行剔除。在地震文本中，主观信息比较少，过滤规则还有待进一步考察。

(2) 分词、词性标注、命名实体识别、句法分析等其他预处理。利用中科院自动化所的分词软件（2006版），对地震文本进行了分词以及词性标注。利用Berkeley Parse[2]句法分析器对过滤后文本中的句子做句法分析。

3.2.2 识别

事件自动抽取首先要对事件词、事件论元以及事件属性等三个重要内容做自动识别。事件词的识别也就是通常说的谓词识别。事件论元识别及事件属性识别和通常意义上所说的语义角色标注(Semantic Role Labeling)类似,其主要任务是分析句子的"谓词—论元"结构,标记出句子中某个谓词出现的论元以及事件属性。

中文Proposition Bank是宾州大学仿照英文PropBank制作的中文语义角色标注语料

[2] 下载地址: http://code.google.com/p/berkeleyparser/downloads/list.

库，主要由两个资源构成：1.语义角色标注语料。2.动词框架。其标记可以分为两级:一级标记是Arg0-4，包含施事、受事、与事等核心论元共5个。二级标记是非核心论元，包括ArgM以及ArgM-ADV，ArgM-TMP等功能性标记。功能标记主要有LOC(地点)、TMP(时间)、MNR(方式)、DIR（方向）、PRP（目的）、CAU（理由）、DIS（语篇）、TPC(话题)、EXT(程度)、NEG（否定）、MOD（情态）、FRQ(频率)，REC（同指）、PRD（次谓词）、ADV（状语）。

我们所说的事件论元和事件属性和PropBank中的两级标记涵盖的成分类似，但是若将EXT(程度)、NEG（否定）、MOD（情态）、FRQ(频率)等和施事、受事、LOC(地点)、TMP(时间)等都标记为语义角色似乎不太合理，因为两者考察的侧重点不同。正因为如此，我们的体系中区分了论元以及事件属性。在下文我们会看到二者的区别。

（1）事件词识别

面对具体的语篇，事件的识别首先转化为事件词的定位。按照三个事件群层级理论，分层次识别事件词。我们将上文中人工标注分析得到的133个事件词作为种子事件词，然后去覆盖地震文本。为了验证事件词识别的效果，我们把同样的200篇地震文本(270kb，43,857字符）作为训练语料，进行封闭测试，另外又选择了50篇地震文本（80.5kb，10,470字符）作为开放测试集进行考察。将机器识别的结果和人工标注结果进行比对，实验结果如下：

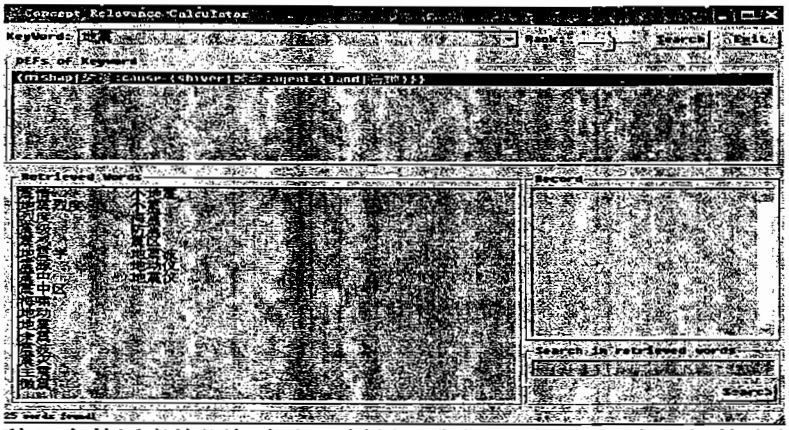
表1：地震类文本三个层面事件词封闭及开放测试实验

地震类文本事件词识别		Precision（%）	Recall（%）	F-Score（%）
本体事件词识别	封闭测试	83.95%	95.21%	89.23%
	开放测试	87.46%	99.27%	92.99%
次生事件词识别	封闭测试	89.10%	95.72%	92.29%
	开放测试	87.13%	96.75%	91.69%
再生事件词识别	封闭测试	88.05%	97.90%	92.72%
	开放测试	81.82%	87.10%	84.38%

在事件词识别的精度上，之所以会出现人工标注和机器识别的误差，一方面和训练语料规模较小有关，另一方面和分词错误有关。比如“【地震】 学家 称 ， 此次 【地震】 与 希腊 西部 周日 发生 的 6.5 级 【地震】 无关 。”“地震学家”被切分为“地震/学家”，导致此处“地震”被错误识别。一些实体的识别还存在问题，导致实体中的成分被标记为事件词，如“在 位于 【震中】 以东 约 240公里 的 首都 雅典 ，【震感】 持续 将近 20秒 。”另外一些副线信息链未完全过滤也造成了识别上的错误。

从上表中来看，因为语料的规模小，导致了封闭测试和开放测试之间数据的差别，尤其是再生事件词的识别率明显下降。克服这个问题的解决方法之一是扩大人工标注语料量的规模，尽可能人工地将三层事件词最大限度地标注提取出来。但是毕竟人工标注是一件辛苦的事情。于是我们想到了利用现有的语义资源HowNet进行相关词的扩充。

在我们购买的知网（2008 版）中，提供三个可执行程序，分别为：Browser (Hownet Knowledge System)、Concept Relevant Calculator 以及 Concept Similarity Measure。其中，在 Concept Relevant Calculator 中，我们能够利用知网内部规定的语义关系、义原联系而直接得到某一个概念的相关词。这种相关词分“Rank Ⅰ”“Rank Ⅱ”“RankⅢ”三个等级。等级增加，检索的相关词也会增加，但是相关度值会降低。下图，显示了“地震”Rank Ⅱ的 Retrieved Words：



利用这种已有的语义资源加上人工判断，我们可以扩充三个层级的事件词，从而在文本内进行事件识别的时候能够尽可能地识别出多的事件词。

（2）事件论元识别

论元是事件的一个重要组成部分。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，文本涵盖的往往是一个典型事件的全部论元在其中的部分映射。例如"地震"类事件关联的论元有"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，在具体句子"甘肃景泰发生一次 Ms4.2 级地震"中，关联论元只有地点"甘肃景泰"和震级"Ms4.2 级"。

论元和语篇结合后，具体表现形式主要有实体表达、时间表达。实体（Entity）是信息提取中的一个基本的概念，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"。利用论元多为实体和时间表达的特征，我们利用句法分析器进行论元的识别。

（3）事件属性识别

事件属性包括事件的模态（Modality）、事件的极性(Polarity)、事件的普遍性（Generality）、事件的时态（Tense）、事件的体态（Aspect）、事件的程度（Extent）、事件的结果（Result）、事件的次数（Frequency）等。事件模态（Modality）是事件发生的可能性，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，一般文本报道的都是已经发生的事情，故而时态一般是"过去"。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。事件程度（Extent）分为"轻微"、"中等"、"强烈"三大类。有的文本在报道事件时会强调事件作用程度，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。

这些信息在自动识别的时候要借助词性和一些特定的规则。比如"极性"确认，一般句子中会出现"没有"、"未"、"尚无"等词语标记，如"俄远东的萨哈林州未有震感，地震也未引发海啸或造成人员及财产损失"。

3.2.3 分类

这里的分类主要是对论元识别的结果进行分类。比如，“4月14日早晨7时49分，青海省玉树藏族自治州玉树县发生7.1级地震。”中“4月14日早晨7时49分”“青海省玉树藏族自治州玉树县”以及“7.1级”是识别出的论元。分类的结果是把三个论元分别和“Arg0时间”“Arg1地点”“Arg2震级”对照。国内对于语义角色分类的研究，Xue Nianwen的研究[1][2]比较系统全面，并得出了一些很有意思的结论。另外，于江德[8]，刘挺等[9]，刘怀军等[10]，丁伟伟等[12][13]等在分类工作上方面提出了很多特征和规则，可以学习借鉴。很明显，这种分类都依赖于分词、“词性标记”、句法分析器的处理结果，需要寻找规则和特征来进行处理。也正因为如此，丰富有效的特征能够帮助提高分类的准确度。在丰富有效句法特征的基础上，我们可以考虑论元角色的义原特征。利用知网，找出承担某一个角色的论元应当具有的义原特色对分类的结果应该有帮助。

4、结语

本文基于资源建设的理念，在分析事件文本信息结构的基础上，对语篇信息抽取整个工作的策略做了一个规划。当前的语义角色识别和分类工作很多都是基于句子的分析研究，对一个完整的语篇实现自动抽取要比句子复杂很多。本文提出的抽取策略是基于一个完整语篇的，即通过主副线信息结构链的构成分析，确立了突发事件文本三级层次的事件框架体系，并在此基础上确立了信息自动抽取的基本策略，即在过滤副线信息链的基础上对事件词识别，进而对论元及事件属性进行识别和分类，形成结构化的数据。

在这个工作的基础上，将来的工作重点将要进一步实体关系和事件关系的架构，解决诸如篇章内的指代消解、旁指等的研究，使其与现有的工作结合起来，从而构建一个完整的语篇信息抽取方案，真正做到让文本的抽取工作准确高效。下一步的工作，我们还需要以这种抽取方案为参照继续研究其他领域文本的抽取特性，以期让信息抽取的工作更加全面。同时，我们希望研究能为自动摘要、机器翻译等自然语言处理的其他领域提供帮助。

参考文献

[1]Nianwen Xue. 2008. Labeling Chinese predicates with semantic roles. Computational Linguistics, 34(2):225-255.

[2]Nianwen Xue and Martha Palmer. 2005. Automatic Semantic Role Labeling for Chinese Verbs, in Proceedings of the 19th International Joint Conference on Artificial Intelligence. Edinburgh, Scotland.

[3] D. Gildea , D. Juraf sky. Automatic labeling of semantic roles [J]. Computational Linguistics , 2002 , 28 (3) :245-288

[4] Thompson CA , Levy R , Manning CD. A generative model for semantic role labeling [C] . In : Proceedings of ECML22003 , LNAI 2837 , Springer Berlin Heidelberg , 2003. 397-408

[5]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.

[6]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.

[7]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.

[8]于江德，樊孝忠，庞文博，余正涛. 基于条件随机场的语义角色标注[J]. 东南大学学报，2007, 23(3)：361-364.

[9]刘挺，车万翔，李生. 基于最大熵分类器的语义角色标注[J]. 软件学报，2007，18(3):565-573.

[10]刘怀军，车万翔，刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报，2007，21(1):79-84.

[11]袁毓林. 语义角色的精细等级及其在信息处理中的应用[J]. 中文信息学报，2007，21(4):10-20.

[12]丁伟伟，常宝宝. 基于最大熵原则的汉语语义角色分类[J]. 中文信息学报，2008，22(6):20-26.

[13]丁伟伟，常宝宝. 基于语义组块分析的汉语语义角色标注[J]. 中文信息学报，2009，23(5):53-61.

[14]汪红林. 基于依存分析的语义角色标注研究[D]. 苏州大学硕士论文，2009.