一种面向突发事件的文本语料自动标注方法

刘炜[1] 王旭[1] 张雨嘉[1] 刘宗田[1]
[1]上海大学计算机工程与科学学院，上海 200444
(wangx89@126.com)

An Automatic-Annotation Method for Emergency Text Corpus

Wei Liu[1] Xu Wang[1] Yujia Zhang[1] Zongtian Liu[1]
[1](School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract Event-based knowledge representation and reasoning has become a research hotspot in Natural Language Processing and Semantic Web. Event-based text corpus is the foundation for the research of event ontology, event-based knowledge reasoning, event-based search, etc. The main purpose of the research on corpus building method is to decrease the ratio of manual labor, and increase the degree of automation on the premise of corpus high quality. Aims at this problem, this paper proposed an event-based text automatic-annotation method for construction of large-scale emergencies news corpus based on multipass filtering, gradually improvement, firstly presented an event structure model, and then, on the basis of LTP analysis, utilized PrefixSpan to mine the rules of event elements from small-scale available corpus, utilized Tonyici Cilin (extended version) to expand the denoters. In the experiment, the automatic-annotation method was compared with manual tagging method and Stanford CoreNLP NER. The experimental results show that this method can improve the efficiency of event-based text annotation effectively.

Key words Emergency Events; Corpus; Automatic-Annotation

摘要 基于事件的知识表示和推理是当前自然语言处理和语义网领域的研究热点，事件语料库是研究包括事件本体、基于事件的知识推理等领域的基础和关键技术之一。对语料库构建方法研究的主要目的是在确保语料库质量的前提下，减少人工参与的比例，增加自动化程度。本文针对该问题，以事件作为文本知识单元，在 LTP 分析的基础上，用序列模式挖掘算法 PrefixSpan 从现有的小规模语料库中挖掘事件要素的词性规则等，用同义词词林（扩展版）对触发词表进行了扩充，采用多遍过滤、逐遍完善的思想提出一种针对大规模突发事件语料库构建的自动标注方法，在实验部分不仅与人工标注做了对比，同时与 Stanford CoreNLP NER 进行了对比，实验效果理想。

关键词 突发事件；语料库；自动标注

中图法分类号 TP391

当前，国内外各类突发事件频发，反映在互联网上则是各类新闻、社交网站关于突发事件的信息呈现爆发式增长。基于大数据的分析思维，这些突发事件文本数据具有巨大潜在的价值，尤其对于舆情系统来说，通过对海量突发事件信息的结构化处理和语义分析实现突发事件的判断和预测具有重要的意义。传统的文本分析手段局限于样本数量和定性研究，无法适应大数据时代对内容挖掘上广度和深度的要求[1]。语料库的分析方法，符合大数据的思维逻辑，通过对海量文本数据的处理，可以对文本内容进行深入挖掘，

收稿日期：
基金项目：国家自然科学基金基于描述逻辑的事件推理关键问题研究(No.61305053)；上海市自然科学基金基于描述逻辑的事件本体形式化表示及推理关键问题研究(No.12ZR1410900)；国家自然科学基金事件本体形式化方法中的几个重要问题(No.61273328)

而不仅仅局限于表层研究或定性分析。

语料库的研究，本质上也是一种跨学科的研究，综合了语言学、修辞学、计算机科学和统计学各学科的知识。通过构建突发事件语料库，可以对突发事件对象进行分析，确定突发事件领域的概念以及概念之间的语义关系，从而可构建针对突发事件的领域本体模型，并进行推理应用。语料库对于实现突发事件领域知识的共享和重用也具有重要意义[2-3]。

语料库建设是自然语言处理技术中的基础性的研究工作，由于事件的特殊性，普通的语料标注方法并不适应于事件标注，因此，学者们对面向事件的语料标注进行了研究。但是限于研究目的和对象的不同，现有事件的语料库分别采用了不同的标注体系[4]。目前，影响较大的事件标注语料库有自动内容抽取（Automatic Content Extraction）评测会议提供的ACE 评测语料[5]、美国高级研究发展学会（Advanced Research and Development Activity）主办的问题回答系统中的时间和事件的识别（Time and Event Recognition for Question Answering Systems）会议的TimeBank 语料[6]。国内在事件标注方面的工作起步较晚，而且缺少大规模的语料库作为研究工作的支撑。更为重要的是，目前大多数的事件语料库是通过手工方式标注，缺点是标注效率低，而且标注过程中人为的主观性容易造成标注标准的不一致，进而影响语料质量。本文在结合上海大学中文突发事件语料库[1]（Chinese Emergency Corpus, CEC）标注规范基础上，提出一种基于事件模型的突发事件语料自动标注方法[2]。

1. 事件模型

本文所标注的文本语料将在文本中标注关于突发事件的完整信息，包括事件的各类要素以及一篇文本中不同事件之间的语义关系。本节简要地介绍事件相关的概念以及事件结构模型。

1.1 事件定义

定义 1（事件）指在某个特定的时间和地点发生的，由若干角色参与，表现出若干动作特征，并伴随着对象内部状态变化的一件事情[7]。对事件的定义可以通过一个形式化的六元组表示：

$$Event ::= < A, O, T, P, S, L >$$

其中 A 表示事件所包含的动作或动作序列的集合，在文本中，动作通常是作为识别一个事件的触发词；O 表示一个事件中的对象集合，包括事件中所有的参与者和涉及到的对象，我们将事件对象分为主体和客体；T 表示事件发生的时间段，事件时间可以是绝对时间也可以是相对时间，两类时间都可以通过计算转换成形如[t1,t2]的序偶表示，以此描述事件的开始、发展和结束时间，当开始时间和结束时间一样时，表示事件发生在瞬间。P 表示事件发生的地点；S 表示事件发生过程中对象的状态集合，由事件发生的前置条件、后置结果集合组成。L 表示事件的语言表现，主要包括事件核心词表现、事件核心词搭配，核心词表现为事件在句子中常用的标志性词汇，通常也是计算机识别事件的触发词，核心词搭配是指核心词与其它词汇的固有搭配。在事件的六个要素中，前五个要素是事件的内在要素。

1.2 事件关系

事件之间的关系分为分类关系和非分类关系。分类关系指事件类之间的包含关系或父子关系，非分类关系指事件或事件类之间内在的语义关系，包括组成关系（isComposedOf）、跟随关系（follow）、因果关系（causal）、并发关系（concurrence）和意念包含关系（thoughtContent）。分类关系通常存在于事件类之间。而在语料标注中，一般只标注非分类关系，因此，以下简单介绍几种非分类关系。

（1）组成关系（isComposedOf）：如果一个大事件 E 可以分解为若干小事件 e_i(其中，i 为正整数)，这些小事件的完成意味着这个大事件完成，则称它们之间具有组成关系，称事件 E_i 是事件 E 的组成部分。例如“雾霾”事件类由“雾霾预警”、“雾霾监控”和“危害评估”等子事件组成。组成关系形式化为 R_{comp}。

（2）因果关系（causal）：事件 e_1 的发生以一定的概率导致了事件 e_2 的发生，发生的概率大于给定的阈值，则称两事件之间存在因果关系，称 e_1 是 e_2 的因，e_2 是 e_1 的果。因果关系形式化为 R_{cause}。因果不但反映了事件之间的相互影响，还在时间上反映了事件发生的先后关系。

（3）跟随关系（follow）：在一定长度的时间段内，事件 e_1 发生后，存在事件 e_2 的接着发生，则称两事件之间存在跟随关系。形式化表示为 R_{follow}。

（4）并发关系（concurrence）：在一定长度的时间段内，存在事件 e_1 和事件 e_2 同时或先后发生（两个事件发生的时间存在重合），则称两个事件之间具有并发关系。形式化为 R_{concur}。

（5）意念包含关系（thoughtContent）：如果一个意念事件 e 包含有若干个意语 e_i，这些意语又分别是多个独立的事件，那么称它们之间的关系是意念包含

[1] https://github.com/shijiebei2009/CEC-Corpus
[2] https://github.com/shijiebei2009/CEC-Automatic-Annotation

关系。

对于一篇文本，通过标注其中的事件、事件要素以及事件之间的关系，形成用以表示文本的事件模型。

2. CEC 及标注规范

2.1 CEC（Chinese Emergency Corpus）

CEC 是前期工作中构建的一个小规模的事件语料库，合计 332 篇，共有五类，分别是地震、火灾、交通事故、恐怖袭击、食物中毒。CEC 与 ACE、TimeBank 语料库相比，规模虽然偏小，但是对事件和事件要素的标注却更加全面。因此，本文将 CEC 作为自动标注研究的训练集与规则挖掘的知识库。

对 CEC 进行分析，其中 Sentences without Event 指不包含事件的句子数目，Event Elements 指事件的所有要素。由表 1 知包含事件的句子占句子总数的 93.48%，触发词占事件所有要素的 41.34%，触发词和事件是一一对应的。

Table 1 Statistics of CEC annotation

表 1 CEC 标注数据统计

Type	Articles	Sentences	Sentences without Event	Events	Denoters	Event Elements
地震	62	401	41	1002	1002	2461
火灾	75	433	39	1216	1216	2935
交通事故	85	514	9	1802	1802	4186
恐怖袭击	49	324	38	823	823	2042
食物中毒	61	392	17	1111	1111	2777
SUM	332	2064	144	5954	5954	14401

定义 2（事件触发词）事件触发词是指在文本中清晰的表示事件发生的词语。

从 CEC 中抽取不同类别的触发词构建触发词表，再用同义词词林扩充触发词表，进而可以用来识别事件。

定义 3（意念事件）一个意念事件是某人心中产生一段意语的事件，这段意语或用口语表达，或用文字描述，或留在心中自知。

定义 4（意念事件触发词）是一个词或词的集合，这些词能够引出意念事件中描述对象内心想法、决策及态度等各方面内容。

意念事件按照动作分类可分为两类：一是诉说类；二是自知类。一段话是一个意念事件，一篇文章是一个意念事件，一个想象是一个意念事件，一个梦也是一个意念事件。如果将意念事件的类型做进一步细分的话，根据对 CEC 的统计可以得到如下分类和举例，见表 2。

Table 2 Categories and examples of thoughts-events denoters

表 2 意念事件触发词分类及举例

分类	例子
narrate(叙述)	称，说，告诉，介绍，表示，谈到，写道，指出，透露……
declare(宣告)	声明，公告，报道，宣传，宣布，公布，发布，告知，宣称，声称……
express(表达)	谴责，质疑，指责，建议，抗议，坦言，强调，解释……
Selfknowing (自知)	希望，期待，愿意，发现，意识到，认为，听说，想到，觉得……
other(其他)	据悉，预见，了解，显示……

定义 5（意语）意语表示行为人用来表达想法、观点、态度和所要描述事实的内容。

简单来说，意念事件触发词所引发的内容即为意语，意语是由意念事件任意一个或共同组成。

Table 3 A sample of annotation text in CEC (partial)

表 3 CEC 标注文本示例（片段）

```
<Body>
  <Title>成都网友称震感强烈 女同事当即哭泣</Title>
  <ReportTime type="absTime">2008 年 05 月 12 日 16:15</ReportTime>
  <Content>
  <Paragraph>
  <Sentence>
    <Event eid="e1">
      <Time type="relTime" tid="t1">5 月 12 日 14 时 28 分</Time>，
       <Location lid="l1">四川</Location>发生 7.8 级
      <Denoter type="emergency" did="d1">地震</Denoter>。
    </Event>
  </Sentence>
  </Paragraph>
  </Content>
<eRelation relType="Causal" cause_eid="e1" effect_eid="e5" />
</Body>
```

2.2 标注规范

CEC 标注的格式采用 XML 语言，在自动标注研究中亦采用 XML 语言来存储标注的语料，各标签的定义以及标签之间的嵌套关系详见图 1。

图 1 中，Denoter 表示事件的触发词，类型共包括七种：突发事件(emergency)、移动事件(movement)、声明类事件(statement)、原子动作事件(action)、操作事件(operation)、状态改变事件(stateChange)、感知事件(perception)；Time 表示时间

要素，其类型包括：相对时间(relTime)、绝对时间(absTime)、段时间(timeInterval)；Location 表示环境要素；Participant 表示事件参与者，其类型包括：主体 Agent、客体 Recipient[9]。其中事件的类型还可以标注为 thoughtEvent，表示意念事件。如果为非意念事件，那么 Event 标签不添加类型属性。Title、ReportTime、Content 及 eRelation 处于并列结构，一个 Content 标签可以包括多个 Paragraph 标签，一个 Paragraph 标签可以包括多个 Sentence 标签，一个 Sentence 标签内可以包括零个或多个 Event 标签。

其中 Event、Denoter、Participant、Time、Location 标签均具有 id 属性，分别为：eid="eN"、did="dN"、sid="sN"、tid="tN"、lid="lN"，属性值中的 N 表示在整篇文章中，其所处的序号。eid 表示事件编号，did 表示触发词编号，sid 表示事件参与者主体的编号，tid 表示时间编号，lid 表示地理环境编号。eRelation 表示事件关系，它的 relType 表示事件关系类型，定义了五种类型的值，分别是：causal、accompany、follow、composite 以及 thoughtContent。表 3 展示的是 CEC 语料标注文本的一个示例片段。

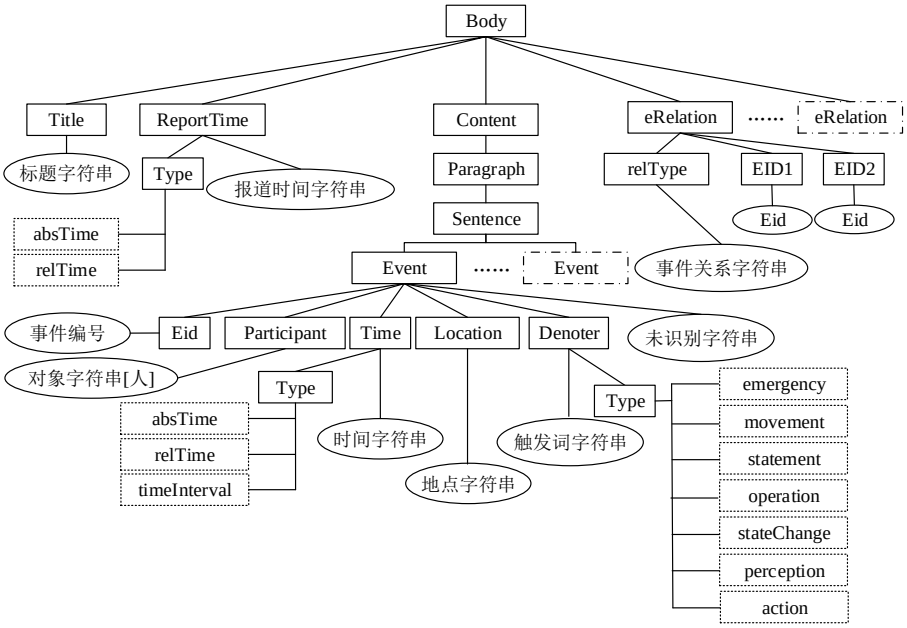


Fig. 1 XML Schema of text automatic-annotation

图 1 自动标注 XML 标签规范

3. 自动标注

实现自动化标注需要多项基础性工作，包括分词、词性识别、命名实体识别、要素识别等。因此，选择合适的分词工具是实现自动化标注的第一步工作。由于一个事件必有一个触发词，我们的方案是借助识别触发词来识别事件，进而识别事件的其它要素，完成基于事件的自动标注。

3.1 分词工具

在现有的分词工具中，LTP（Language Technology Platform）[10]制定了基于 XML 的语言处理结果表示，并在此基础上提供了一整套自底向上的丰富、高效、高精度的中文自然语言处理模块，以及能够以网络服务使用的语言技术云，因此选用 LTP 作为分词工具。

例如，图 2 是采用 LTP 对“*2014 年 10 月 7 日晚 21:49:39，在云南省普洱市景谷县发生 6.6 级地震。*”进行分词、词性标注以及命名实体识别分析后的结果。其中第一行表示分词的内容，第二行表示分词内容的 id 号，第三行表示词性标注，第四行表示命名实体识别，第五行表示依存句法分析（其中 ATT 表示定中关系，ADV 表示状中结构，COO 表示并列关系，POB 表示介宾关系），第六行表示父节点的 id 号，第七行表示语义角色标注。

LTP 词性标注采用 863 词性标注集，命名实体识别模块采用 O-S-B-I-E 标注形式，其中 O 表示这个词不是 NE（Named Entity），S 表示这个词单独构成一个 NE，B 表示这个词为一个 NE 的开始，I 表示这个词为一个 NE 的中间，E 表示这个词为一个 NE 的结尾；核心的语义角色为 A0-A5，A0 通常是动作的施事，A1 通常表示动作的影响，A2-A5 根据谓语动词不同含义不同；LTP 中的 NE 模块可以识别三种 NE，分别是：Nh 表示人名、Ni 表示机构名、Ns 表示地名。其余的语义角色为附加语义角色，如 LOC 表示地点，TMP 表示时间等。

2014年	10月	7日	晚	21	:	49	:	39	，	在	云南省	普洱市	景谷县	发生	6.6	级	地震	。
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
nt	nt	nt	nt	m	wq	m	wq	m	wq	p	ns	ns	ns	v	m	q	n	wq
O	O	O	O	O	O	O	O	O	O	O	B-Ns	I-Ns	E-Ns	O	O	O	O	O
ATT	ATT	ATT	ATT	ADV	WP	COO	WP	COO	WP	ADV	ATT	ATT	POB	HED	ATT	ATT	VOB	WP
1	3	3	4	14	6	4	8	4	8	14	12	13	10	-1	16	17	14	14
											|----------LOC----------|				|---A0----|		

Fig. 2 Analysis results by LTP

图 2 LTP 分析结果

3.2 识别触发词（Denoter）

图 3 为 CEC 中触发词的统计结果，使用 LTP 对 CEC 所使用的生语料(即标注前的原文本)进行分析，可以获得详细的分词与词性标注信息，称之为 Doc-LTP，将 CEC 中人工标注的文本称为 Doc-CEC。针对每一个文本文件进行处理，将 Doc-LTP 与 Doc-CEC 中的同一篇文本进行比较，找到 Doc-CEC 标注出的 Denoter 内容在 Doc-LTP 中所对应的词性，经过对 332 篇语料进行统计，得到触发词的词性是动词、名词（或者包含动词、名词）的次数是 5548 次，触发词总计出现 5896 次。由此得到动词、名词或者包含动词、名词的组合形式在总的触发词词性中所占的比例为 94.0976%。在自动标注时，可以基于构建的触发词表以及统计得到的触发词词性规律来识别触发词。

统计触发词词性算法描述如下：

Step1：将 CEC 语料所使用的生语料抽取出来，即将人工标注过程中所添加的各种标签去掉，还原为未经过任何处理的状态，记为 RC（Raw Corpus）；

Step2：对 RC 进行遍历，得到一篇生语料，记 RC_i；

Step3：用 LTP 对 RC_i 进行分析，得到分词、id 号、词性标注、命名实体识别、语义角色标注信息等；

Step4：将 LTP 分析的结果存入<Key,Value>键值对的集合中，所有的键值对集合的 Key 都是 id 号，这样能够根据分词内容获取到其对应的词性等信息；

Step5：开始解析 RC_i 所对应的 CEC 中经过人工标注之后的同一篇语料，取得<Denoter>标签中所标注的所有内容，记为 TW（Trigger Words）；

Step6：对 TW 进行遍历，记为 TW_i，与 LTP 分词的结果进行比较，得到与触发词内容相同的分词串；

Step7：得到分词串所对应的 id 号，根据 id 号查找<id,pos>（pos 表示词性标注）键值对，获得 id 号对应的词性标注结果。

3.3 扩充触发词表

由于 CEC 语料库规模有限，从中抽取的触发词所构建的触发词表亦必然有限，难以做到大规模的覆盖度。因此，本文使用《同义词词林（扩展版）》[11]来扩充触发词表。通过获得触发词的同义词扩展出新的触发词，如“出生”的同义词为：

诞生 出生 降生 生 落地 坠地 出世

上述词都是表示出生这一含义的词，在文本标注过程中这些词通常会触发一个“出生”类事件。

扩充触发词表算法描述如下：

Step1：对某一类触发词表，遍历触发词表中的每一个词 W_i，在同义词词林中查出它的全部同义词项；

Step2：取该词所在的同义词项的总词数为 S；

Step3：统计该词项中其他的词汇出现在该类触发词表中的个数为 N（包括 W_i 自身）；

Step4：计算 N/S，如果 $N/S \in [0.4,1]$，本次实验下限阈值取为 0.4；

Step5：那么取出这个义项中所有不在当前触发词表中的词汇，并且计算该词汇的长度，以便识别是单字还是词汇；

Step6：将属于词汇的同义词项全部扩展到触发词表中（舍弃单字的同义词项）。同样的，使用该方法扩展其他类别的触发词表；

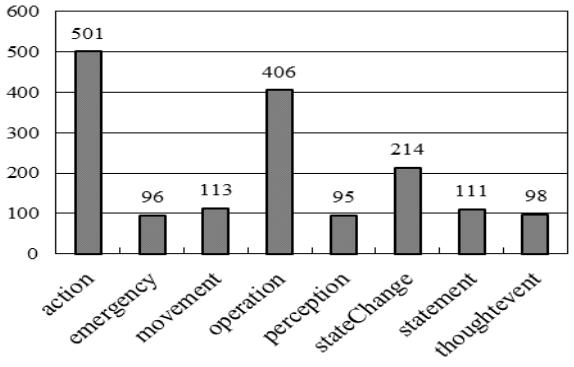


Fig. 3 Statistics of 8 categories of denoters in CEC

图 3 CEC 语料中 8 类触发词数量统计图

图 3 显示的是从 CEC 语料中提取出的 8 类触发词统计结果图，经过扩充后得到的触发词表分类统计如图 4。

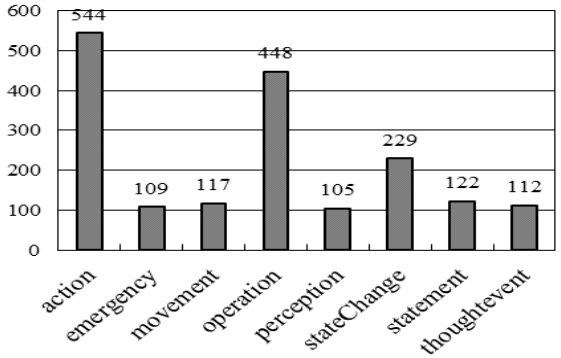


Fig. 4 Statistics of denoters after expansion

图 4 扩充后触发词数量统计图

3.4 识别 Participant、Location、Time 要素

同样的，使用识别 Denoter 的方法，还可以从 CEC 中抽取出 Participant、Location、Time 要素所对应的词性集合，对于抽取出来的词性集合，每一个要素内容所对应的词性规则是有序且可重复的。例如，一个 Location 要素内容的词性规则是：[ns,nd,ns,ns,nd]，之后使用序列模式挖掘算法从大量的词性规则中挖掘频繁序列，对于挖掘的结果要进行人工筛选，并添加一些人工构建的规则，序列模式挖掘算法采用文献[12]提出的 PrefixSpan 算法，虽然文本内容的形式会多种多样，但是不同的文本其词性是固定的。因此，构建基于词性的识别方法是可以应付文本内容多样化的情况的。限于篇幅，仅列举几例作为说明：

例 1：“*当地时间 7 日凌晨 1 点 45 分左右*”

LTP 分词及词性标注：“当地 nl 时间 n 7 日 nt 凌晨 nt 1 点 nt 45 分 nt 左右 m”，在识别时间要素时，可以从开始的 nt 节点一直扫描到连续的最后一个 nt 节点，即 nt+，（“+”表示出现 1 次或多次，“*”表示出现 0 次或多次，“？”表示出现一次或一次也没有，“|”表示或者，“&”表示并且，“->”表示紧跟）将其作为 Time 要素。

例 2：“*当地时间 16 时 30 分*”

LTP 分词及词性标注：“当地/nl 时间/n 16 时/nt 30 分/nt”

例 3：“*当地时间 16:30，我们出发了*”

LTP 分词及词性标注：“当地/nl 时间/n 16:30/m ，/wp 我们/r 出发/v 了/u”，由 nl 开始紧跟 n 直到连续的最后一个 nt 或 m 结束，即 nl(n)(m|nt+)识别 Time 要素。

例 4：“*中国国家主席习近平、国务院总理李克强*”

LTP 返回的 XML 格式标注结果：

```
<word id="0" cont="中国" pos="ns" ne="S-Ns" parent="2" relate="ATT" />
<word id="1" cont="国家" pos="n" ne="O" parent="2" relate="ATT" />
<word id="2" cont="主席" pos="n" ne="O" parent="3" relate="ATT" />
<word id="3" cont="习近平" pos="nh" ne="S-Nh" parent="-1" relate="HED"/>
<word id="4" cont="、" pos="wp" ne="O" parent="7" relate="WP" />
<word id="5" cont="国务院" pos="ni" ne="S-Ni" parent="6" relate="ATT" />
<word id="6" cont="总理" pos="n" ne="O" parent="7" relate="ATT" />
<word id="7" cont="李克强" pos="nh" ne="S-Nh" parent="3" relate="COO"/>
```

由上例得出，使用 S-Ns+(S-Nh+)S-Ni?S-Nh+可以识别 Participant 要素。

例 5：“*云南省昆明市石林彝族自治县境内*”

LTP 返回的 XML 格式标注结果：

```
<word id="0" cont="云南省" pos="ns" ne="B-Ns" parent="1" relate="ATT"/>
<word id="1" cont="昆明市" pos="ns" ne="I-Ns" parent="2" relate="ATT" />
<word id="2" cont="石林" pos="ns" ne="I-Ns" parent="4" relate="ATT" />
<word id="3" cont="彝族" pos="nz" ne="I-Ns" parent="4" relate="ATT" />
<word id="4" cont="自治县" pos="n" ne="E-Ns" parent="5" relate="ATT" />
<word id="5" cont="境内" pos="nl" ne="O" parent="6" relate="ADV" />
```

根据 LTP 的命名实体的标识说明，我们用 B-Ns(I-Ns*)E-Ns(nl?|nd?)识别 Location 要素。

对所挖掘的词性规则以及人工构建的规则进行汇总如表 4 所示。

Table 4 rules for identification of event elements

表 4 事件要素识别规则

要素	词性规则	语义角色
Time	(nt+)m(wp)m	TMP+(DIS?)
	nl(n)(m| nt+)	
	nt+(m)q	
	nt+(m?)	
要素	词性规则&依存句法分析	命名实体
Participant	m(n)& ATT(VOB|SBV|DBL)	S-Ni+(S-Nh)
	m(q)n&ATT(ATT)(VOB|SBV|DBL)	S-Ns+(S-Nh)
		B-Ni(I-Ni*)E-Ni(S-Nh)
		S-Nh+(S-Ni?)
		B-Nh(I-Nh*)E-Nh
要素	命名实体->词性规则	语义角色
Location	B-Ns(I-Ns*)E-Ns->(nl?|nd?)	LOC+
	B-Ns(I-Ns*)E-Ns(S-Ns+)	
	(S-Ns+)->(nl?|nd?)	
	(S-Ns+)	

上述各列均可以作为独立的识别规则。

3.5 多遍过滤的自动标注方法

在自动标注过程中，一遍标注很难识别出所有的要素以及事件的边界，而采用多遍过滤的方法可以对文本标注的结果不断修正和逐步完善。图 5 所示为自动标注的流程图。以下对其中主要的步骤进行详细的说明，次要的步骤简略说明。

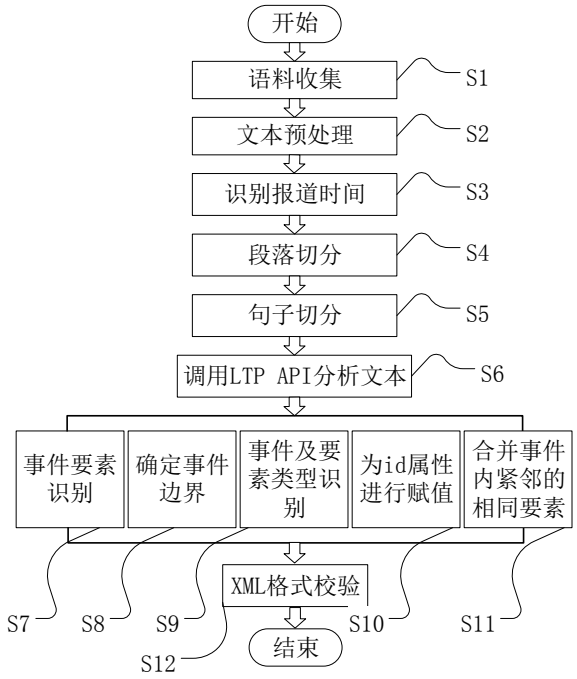


Fig. 5 Flow chart of automatic-annotation

图 5 自动标注方法流程图

S1：收集生语料：从互联网上收集关于突发事件的新闻报道作为生语料，包括地震、火灾、交通事故、恐怖袭击及食物中毒。

S2：预处理步骤将新闻报道的时间放在第一行，之后为新闻报道的内容。

S3：为了判断时间的类型，构造正则表达式如下“^.*(\\d{4}(.)+\\d{1,}(.)*\\d{1,}(.)*)|(\\d{2}年\\d{1,}(.)*\\d{1,}(.)*)$”如果匹配，是绝对时间，否则是相对时间。

S4：将报道内容按段落切分。

S5：以“。！？；”为分割符，切分句子。

S6：将每一个独立的完整的句子作为待分析文本，调用 LTP 的 API 分析文本，设置返回结果格式为 XML。

S7：根据构建的触发词表以及挖掘的词性规则来识别事件的要素。

S8：扫描经过第一遍过滤处理之后的文本，当发现<Event>标签之后，作一个标记，继续扫描，如果发现另一个<Event>标签或者</Sentence>标签，则在这些标签之前插入</Event>标签，以此确定事件的边界。

S9：扫描经过第二遍过滤之后的文本，根据之前构建的八种类型的触发词表对触发词的属性进行识别，并添加对应的属性信息，如果触发词在意念事件的触发词表中，要对<Event>标签添加 type="thoughtevent"属性，将触发词的类型设置为 type="statement"；对于 Time 要素，使用 S3 步骤中的正则表达式来对 Time 的时间类型进行识别。

S10：扫描经过第三遍过滤之后的文本，设置一个计数器，从 1 开始，当扫描到一个标签之后，开始为其 id 属性进行赋值，如果不是<Event>标签，那么在<Event>标签之内的所有标签的 id 属性的值和<Event>标签 id 属性值相同，当扫描到下一个<Event>标签时将计数器加 1 再进行赋值。

S11：扫描经过第四遍过滤之后的文本，对于一个<Event>标签之内，如果有两个紧邻的相同标签或者两个相同的标签之间的字符不超过 2，那么合并相同的标签以及标签内的内容。

S12：本文采用 DTD 文件对 XML 文件进行格式校验。

上述步骤中，S11 中合并紧邻相同要素的标签及内容算法如下：

Step1：扫描经过四遍过滤之后的文本，从中抽取出所有的 Event 节点，记作 EN；

Step2：遍历每一个 Event 节点 EN_i，抽取出 EN_i 内所有的子节点，记作 CN；

Step3：遍历每一个 CN 子节点 CN_i，获取 CN_i 标签的名称，记作 priorNodeName，获取 CN_i 标签内的值，记作 priorNodeValue；

Step4：继续遍历 CN 子节点的 $CN_{(i+1)}$ 节点，获取 $CN_{(i+1)}$ 标签的名称，记作 nextNodeName,获取 $CN_{(i+1)}$ 标签内的值，记作 nextNodeValue；

Step5：获取 CN_i 节点和 $CN_{(i+1)}$ 节点之间的值，记作 median，得到 median 的长度，记作 median_len；

Step6：如果 priorNodeName==nextNodeName，并且 median_len<=2，将 priorNodeValue、median、nextNodeValue 进行拼接得到最终的值 final_value；

Step7：删除 median，删除 $CN_{(i+1)}$ 节点，用 final_value 代替 CN_i 节点的 priorNodeValue 值；

Step8：从头开始重新循环遍历。

4. 实验与分析

4.1 实验 1—要素识别

目前，准确率、召回率和 F_1 值是被广泛用来评价检索系统实验结果的标准。召回率衡量的是检索结果的查全率；准确率衡量的是检索结果的查准率。F_1 值则结合了准确率和召回率，是评价检索结果的最常用方法。本文用这三个标准来评价自动标注的效果。采用 CEC 作为实验数据，使用程序自动标注之后将其与人工标注语料进行详细的对比。

由于研究的目的在于实现自动标注，而不是进行精确的文本匹配。所以在实现过程中，更侧重于要素的识别。例如，人工标注过程中将“当地时间 1 月 14 日晚”识别为 Time 要素，而在自动标注中可能会

将“当地时间 1 月 14 日”或者“当地时间 1 月 14 日晚，”（含标点符号）识别为 Time 要素。在实验过程中，认为这两种自动标注情况都是正确的。

定义自动标注识别正确个数为 E_r，自动标注识别总个数为 E_t，人工标注识别总个数为 E_a，准确率、召回率、F_1 值的计算方法如下所示：

准确率（P）：$P=\frac{E_r}{E_t}$ 召回率（R）：$R=\frac{E_r}{(E_a+E_t)/2}$

F_1 值（F_1）：$F_1=\frac{2\times P\times R}{P+R}$

在实验过程中，由于没有权威的对比语料以及评价方法，暂且认为人工标注的准确率已足够高。但是未必达到百分之百，所以在计算召回率的时候，首先计算了自动标注识别个数与人工标注识别个数的平均值作为分母，这样在没有标准对比实验语料的情况下，既考虑到了自动识别也兼顾了人工识别。经过对 CEC 的实验，标注要素个数统计如表 5 所示，实验结果如表 6 所示。

Table 5 Statistics of element annotation

表 5 要素标注统计

要素	自动标注识别正确的个数	自动标注识别的总个数	人工标注识别的总个数
Denoter	5583	7523	4937
Time	1320	1935	1329
Location	1035	1607	1476
Participant	1195	1802	2928
ReportTime	314	332	332

Table 6 Experimental results

表 6 实验结果

要素	准确率	召回率	F1 值
Denoter	74.21%	89.62%	81.19%
Time	68.22%	80.88%	74.01%
Location	64.41%	67.14%	65.75%
Participant	57.36%	50.53%	57.36%
ReportTime	94.58%	94.58%	94.58%

4.2 实验 2—事件识别

对 CEC 人工标注的语料进行统计，发现共标注了 5954 个事件，使用程序对 332 篇生语料完成自动标注之后，统计显示共标注了 7523 个事件，如表 7 所示。从数量上来看，使用程序标注出的事件多于人工标注出的事件。这是因为相对于人工来说，程序实现的自动标注都是基于分词工具的分词结果，而分词工具都是较细粒度的对字词进行切分。事件触发词是指在文本中清晰的表示事件发生的词语，自动标注在识别触发词之后会基于一个事件必有一个触发词的原则，认为这个触发词一定是属于某个事件的，而事件的其他要素是可以缺省的，从而导致了自动标注的事件数量比人工标注的事件数量多。这也是本方法的不足之处，在后期需要改进的地方。

Table 7 Comparisons of event identification

表 7 事件识别对比

人工标注事件个数	5954
自动标注事件个数	7523

4.3 实验 3—与 Stanford Named Entity Recognizer (NER)识别对比

为了更客观的对本文方法进行说明，采用 Stanford Named Entity Recognizer (NER) [13]进行对比实验。NER 模块提供了针对 English、German、Chinese 的分类器，并分别提供了均衡语料和非均衡语料。NER 使用的训练数据有两类，分别是 CTB（美国宾州大学构建的中文树库）和 PKU（中国北京大学构建的树库）。对于英文，NER 可以根据不同的训练语料分别提供三类（Location、Person、Organization）模型、四类（Location、Person、Organization、MISC）模型和七类（Time, Location, Organization, Person, Money, Percent, Date）模型。但是对于中文，NER 要求输入集是中文分词的输出集，并且仅识别 GPE(Geo-Political Entity)、PERSON、LOC(Location)、ORG(Organization)、MISC(Names of Miscellaneous Entities)，可以看出 MISC 作为杂项结果集，也就是不能够准确识别为某一种具体的 NER 集合。

使用 NER 对 CEC 中 332 篇生语料进行识别，为了实验的公正性，分别采用了 CTB 结合非均衡语料和 PKU 结合非均衡语料进行实验，以此与自动标注识别的要素进行对比。

基于上面的说明，在本文的事件自动标注过程中，Participant 要素对应 ORG 和 PERSON，Location 要素对应 LOC 和 GPE，因为 MISC 是杂项结果集，所以将其分别与 Participant 和 Location 进行对比，但是任一个识别项只会择 Participant 或 Location 其一，不会出现同时匹配两者的情况。从 NER 标注过的同一篇文本中，统计与自动标注的语料有交集的数目，对 332 篇语料汇总之后，实验结果如下:

Table 8 Comparisons with CTB unbalanced identification

表 8 CTB 非均衡语料识别对比

交集名称	数目	要素名称	数目	交集占比
ORG∩Participant	442	Participant	1802	88.40%
PERSON∩Participant	246			
MISC∩Participant	905			
LOC∩Location	128	Location	1607	89.61%
GPE∩Location	1110			
MISC∩Location	202			

Table 9 Comparisons with PKU unbalanced identification

表 9 PKU 非均衡语料识别对比

交集名称	数目	要素名称	数目	交集占比
ORG∩Participant	440	Participant	1802	88.51%
PERSON∩Participant	238			
MISC∩Participant	917			
LOC∩Location	137	Location	1607	89.61%
GPE∩Location	1111			
MISC∩Location	192			

由实验结果可以看出，自动标注方法识别的要素在 NER 中同样被识别或者说 NER 识别的实体中有部分可被自动标注方法的 Participant 和 Location 要素所识别，同时两者所共同识别或者有交集部分对自动标注识别的要素的覆盖度在 88%以上。实验结果说明基于挖掘的规则以及 LTP 标注出的命名实体识别 Participant 和 Location 要素正确率较高。

5. 结束语

突发事件语料库在对海量 Web 信息进行基于事件的文本分析中发挥了重要的作用，是实现网络突发事件自动判断和预警的重要基础工作之一，同时也是构建事件本体实现事件知识共享的重要基础。本文针对现有手工构建事件语料库的不足，提出一种新的语料自动标注方法。通过实验表明，对于新闻报道类的文本，本文所提出的方法能够有效地对生语料进行自动化的标注，提高了语料标注的效率。相比于传统的人工标注方法具有以下优点。

（1）该方法采用程序实现自动标注，可以极大的提高标注速度。

（2）在识别准确率不高的情况下，可以作为人工标注的前期工作。用程序自动标注之后，人工对部分内容做调整，非常有利于大规模的语料标注工作。

（3）对标注后的 XML 内容进行格式检查，确保自动标注语料的质量，同时标注格式满足中文突发事件语料库规范。

（4）采用多遍过滤的思想，便于后期对识别方法进行改进，一旦有更好的识别方法，可以将其加入到过滤链条之中。

本文的方法仍存在需要改进的地方，主要体现在目前触发词和事件要素的自动识别准确度还不能达到非常理想的程度，另外事件关系的识别及推理还需要深入研究。

参 考 文 献

[1] 喻国明, 李慧娟. 大数据时代传播研究中语料库分析方法的价值[J]. 传媒, 2014 (2): 64-66.

[2] LI Xiang, LIU Gang, LING An-hong, et al. Building a practical ontology for emergency response systems[C]//Proceedings of 2008 International Conference on Computer Science and Software Engineering. 2008: 222−225

[3] Q YU Kai, WANG Qing-quan, RONG Li-li. Emergency ontology construction in emergency decision support system[C]//Proceedings of 2008 IEEE International Conference on Service Operations and Logistics, and Informatics. 2008: 801−805

[4] 付剑锋. 面向事件的知识处理研究[D]. 上海: 上海大学, 2010

[5] Doddington G R, Mitchell A, Przybocki M A, et al. The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation[C]//LREC. 2004

[6] Pustejovsky J., Hanks P., Sauri R., et al., The timebank corpus [EB]. In Corpus Linguistics, 2003, pp.647~656, http://ucrel.lancs.ac.uk/publications/cl2003/papers/pustejovsky.pdf

[7] 刘宗田, 黄美丽, 周文, 等. 面向事件的本体研究[J]. 计算机科学, 2009, 36(11): 189-192

[8] Consortium, L.D., 2005. ACE（Automatic Content Extraction）chinese annotation guidelines for events. http://projects.ldc.upenn.edu/ace/docs/Chinese-Entities-Guidelines_v5.5.pdf

[9] Zhang X, Liu Z, Liu W, et al. Research on event-based semantic annotation of Chinese[C]//Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on. IEEE, 2012: 1883-1888.

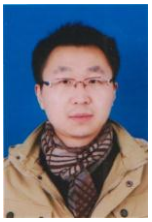
[10] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China

[11] http://www.ltp-cloud.com/download/#down_cilin

[12] Pei J, Han J, Mortazavi-Asl B, et al. Mining sequential patterns by pattern-growth: The prefixspan approach[J]. Knowledge and Data Engineering, IEEE Transactions on, 2004, 16(11): 1424-1440

[13] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363-370

Liu Wei was born in 1978. Associate professor in Shanghai University. He received his Ph.D. degree from Shanghai University in 2005. His research interests include NLP, semantic ontology technologies and knowledge representation.



Wang Xu was born in 1989. He is master candidate in Shanghai University. His main research interests include knowledge representation and machine learning.

Zhang Yujia was born in 1992. She is master candidate in Shanghai University. Her main research interests include knowledge representation and machine learning.

Liu Zongtian was born in 1946. He is professor and Ph.D. supervisor in Shanghai University. His main research interests include artificial intelligence and software engineering.