

文章编号: 1003-0077(2017)00-0076-10

一种面向突发事件的文本语料自动标注方法

刘 炜,王 旭,张雨嘉,刘宗田

(上海大学 计算机工程与科学学院,上海 200444)

摘 要: 事件语料库是研究语义 Web 中事件知识的抽取、表示、推理和挖掘的基础和关键技术之一。该文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,结合自定义的事件要素词典,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

中图分类号: TP391 **文献标识码:** A

An Automatic-Annotation Method for Emergency Text Corpus

LIU Wei, WANG Xu, ZHANG Yujia, LIU Zongtian

(School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: Event-based text corpus is the foundation for the research on detection, representation, reasoning and exploitation of events in the Semantic Web. This paper proposes an automatic-annotation method for event-based texts to construct large-scale emergencies news corpus. Firstly, this paper presents an event structure model as event-based knowledge unit; Secondly, on the basis of text process by LTP, we apply the PrefixSpan to mine the rules of event elements from small-scale available corpus; Thirdly, by combining a customized dictionary of event elements, the denoters are expanded by Tonyici Cilin (Extended). In the experiment, the automatic annotation method is compared with manual tagging method and Stanford CoreNLP NER, showing that this method can improve the efficiency of event-based text annotation effectively.

Key words: emergency events; corpus; automatic; annotation

1 引言

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。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在内容挖掘上对广度和深度的要求^[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,而不仅仅局限于表层研究或定性分析。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^[2-4]。

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, 现有的事件语料库分别采用了不同的标注体系^[5]。这些标注体系主要关注某些特定类型的事件或事件要素, 忽略了一般意义上的事件以及人们对于事件的理解和认识。目前, 影响较大的事件标注语料库有 ACE 评测语料^[6-7]和 TimeBank 语料^[8]。其中 ACE 的评测任务只针对特定类型的事件及其子类事件, 因此语料中也只标注了这些特定类型的事件信息, 除了事件的类型和子类型之外, ACE 中的事件还具有四种属性: 事件的极性、事件的时态、事件的指属、事件的形态; TimeBank 标注了事件、时间、时间指示词以及事件和时间之间的关联关系等等, 另外其采用了一种改进的 XML 语言-TimeML 进行标注, 增强了它在描述时间信息方面的能力。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。主要的工作有上海大学的中文突发事件语料库^① (Chinese Emergency Corpus, CEC)。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更为全面, 详见表 1 的对比分析。纵观现有大部分事件语料库, 多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 CEC 语料库标注规范基础上, 提出一种基于事件模型的突发事件语料自动标注方法^②。

表 1 CEC 与 ACE 和 TimeBank 对比

	CEC	ACE	TimeBank
支持语言	中文	中文/英文	英文
文本篇数	332	633/4090	300
事件数	5954	2521/4090	7571
标注所有事件	YES	NO	YES
标注事件要素	YES	YES	YES
标注事件关系	YES	NO	YES
标注模式	基于语义	基于事件	基于动词

2 事件模型

在自然语言处理领域, “事件”是一个非常重要的概念。事件关系到多方面的静态概念, 是比静态概念粒度更大的知识单元。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。

2.1 事件定义

定义 1 事件 (Event), 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^[9]。对事件的定义可以通过一个形式化的六元组表示, 如式 (1) 所示。

$$e::=_{def}(A,O,T,V,P,L) \tag{1}$$

A 表示动作; O 表示对象; T 表示时间; V 表示地点; P 表示断言; L 表示语言表现。

定义 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 (concurrency) 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。关于事件和事件关系的语义定义见文献^[9]。

3 CEC 及标注规范

3.1 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。语料文本分为五类, 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更为全面。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 2 可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%, 触发词和事件为一一对应。

定义 3 事件触发词, 指在文本中清晰地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。

① <https://github.com/daselab/CEC-Corpus>
② <https://github.com/daselab/CEC-Automatic-Annotation>

表 2 CEC 标注数据统计

Type	Articles	Sentences	Sentences without Event	Events	Denoters	Event Elements
地震	62	401	41	1 002	1 002	2 461
火灾	75	433	39	1 216	1 216	2 935
交通事故	85	514	9	1 802	1 802	4 186
恐怖袭击	49	324	38	823	823	2 042
食物中毒	61	392	17	1 111	1 111	2 777
SUM	332	2 064	144	5 954	5 954	14 401

定义 4 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 5 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(表 3)。

表 3 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……
self-knowing(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 6 意语，表示行为人来表达想法、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。

3.2 标注规范

CEC 标注的格式采用 XML 语言，在自动标注研究中沿用 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(move-ment)、声明类事件(statement)、原子动作事件(action)、操作事件(operation)、状态改变事件(state-Change)、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 表示地点要素；Participant 表示事件参与者，其类型包括：主体 Agent、客体 Recipient^[10]。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"、oid="oN"、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，oid 表示事件参与者客体的编号，tid 表示时间编号，lid 表示地点编号。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(因果)、accompany(伴随)、follow(跟随)、composite(组成)以及 thoughtContent(意念包含)。

3.3 标注质量保证

标注语料采用 XML 格式进行存储，可以通过 DTD 或者 XML Schema 对 XML 文件的结构以及嵌套要素进行校验，如果一篇语料有多个不同的标注版本，则计算其一致性，如式(2)所示。

$$agreement = n \frac{|A_1 \cap A_2 \cap \dots \cap A_n|}{|A_1| + |A_2| + \dots + |A_n|} \quad (2)$$

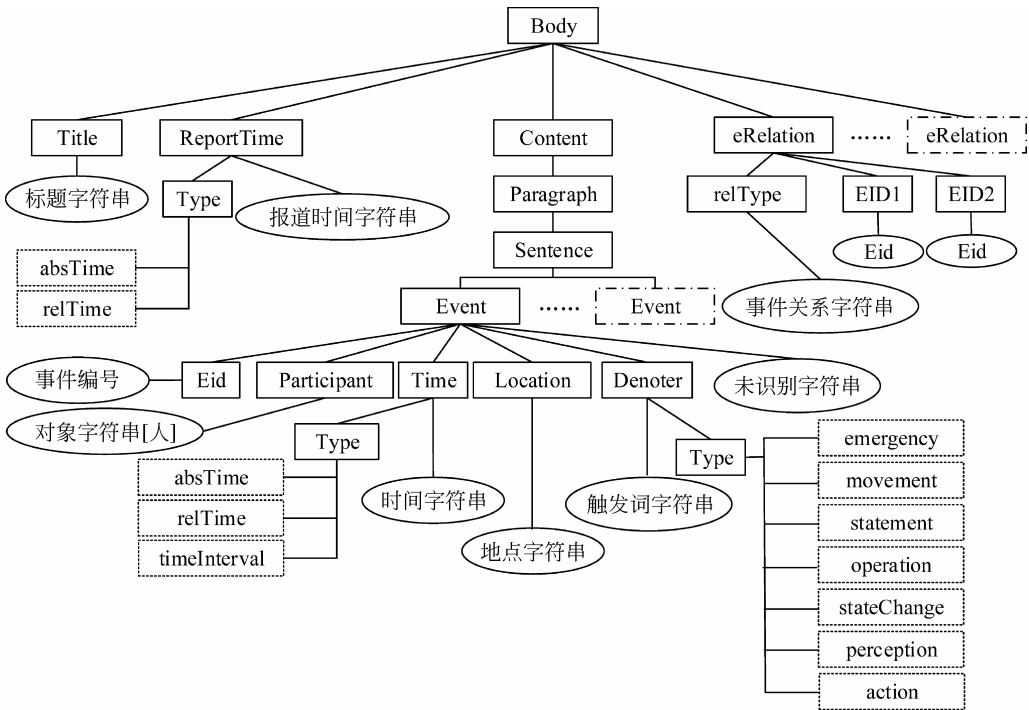


图 1 自动标注 XML 标签规范

$|A_1|$ 表示语料 A_1 中被标注为事件指示词及事件要素的词个数， $|A_1 \cap A_2 \cap \dots \cap A_n|$ 表示 n 种标注版本中标注相同的词的个数。如果 agreement 大于指定的阈值，则将该语料加入语料库中，完成标注，否则，该语料分歧性太大，重新标注，直至其一致性大于指定的阈值。

4 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^[11]。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。

4.1 分词工具

在现有的分词工具中，LTP(Language Technology Platform)^[12] 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。

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。

4.2 识别触发词(Denoter)

图 2 是从 CEC 语料中提取出的八类触发词统计结果图，使用 LTP 对 CEC 所使用的生语料(未标注文本)进行分析，可以获得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。针对每一篇文本文件进行处理，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 5 548 次，占有触发词的比例为 94.097 6%。因此，在自动标注时可基于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：

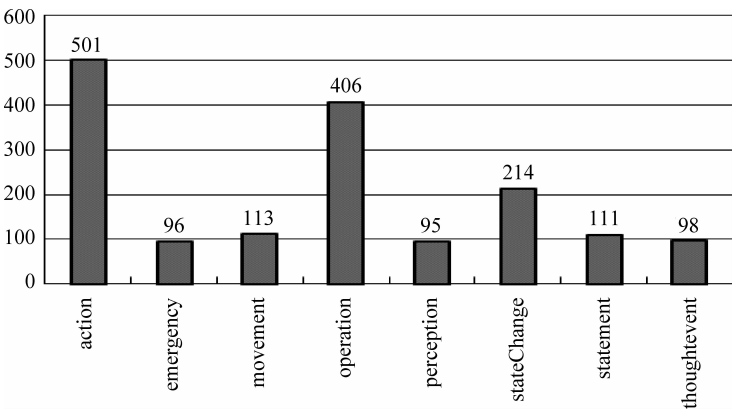


图 2 CEC 语料中八类触发词数量统计图

- Step1: 将 CEC 语料进行去标签处理,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 $\langle \text{Key}, \text{Value} \rangle$ 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 $\langle \text{Denoter} \rangle$ 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 $\langle \text{id}, \text{pos} \rangle$ (pos 表示词性标注)键值对,获得 id 号对应的词性标注结果。

4.3 扩充触发词表

由于 CEC 语料库规模有限,构建的触发词表规模必然有限,难以做到大规模的覆盖度。本文使用

《同义词词林(扩展版)》^[13]来扩充触发词表。如触发词“出生”可扩展为:

诞生 出生 降生 生 落地 坠地 出世
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;

经过扩充后得到的触发词表分类统计如图 3。

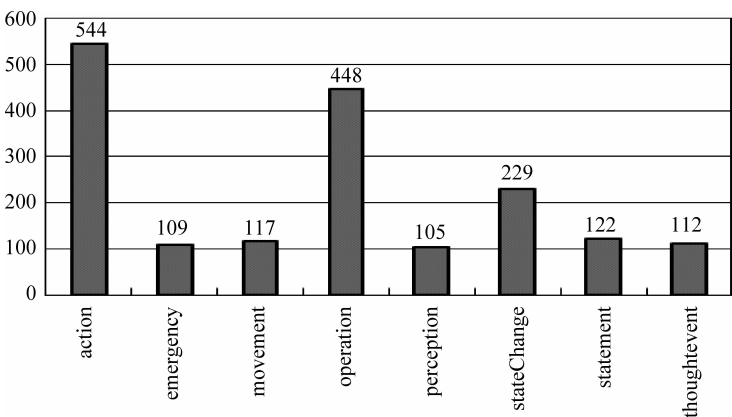


图 3 扩充后触发词数量统计图

4.4 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[14]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,我们出发了”

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,/wp 我们/r 出发/v 了/u”,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 “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 3 “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。

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。

4.5 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 4 所示为自动标注的流程图。

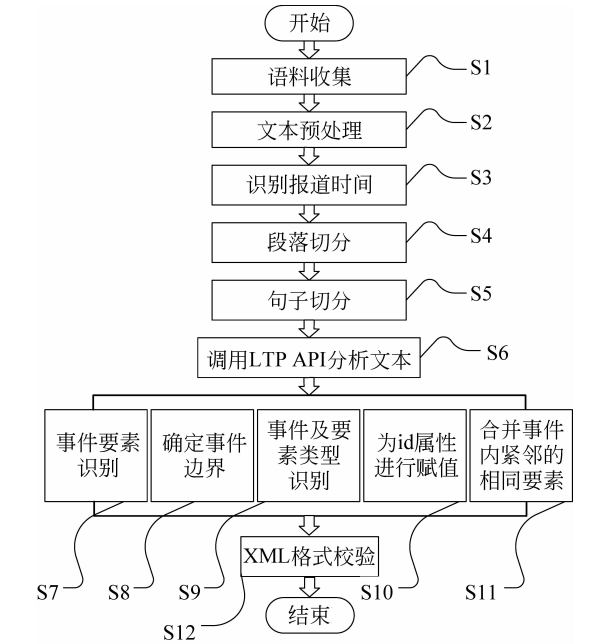


图 4 自动标注方法流程图

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。

- 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,构造正则表达式对时间进行判断是绝对时间还是相对时间;
- 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;之后根据用户构建的自定义事件要素词典再次对事件要素进行识别;
- 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;

续表

- 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。

用户自定义事件要素词典可以收录特殊行业或者自动标注中难以识别的文本,而这些文本是人工可以认定的确是事件某要素的情况,经过多次的迭代,自定义事件要素词典得到不断的扩充与完善,使得自动标注的准确率进一步提高。

5 实验与分析

5.1 实验 1—要素识别

本文通过准确率、召回率和 F_1 值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”识别为时间要素,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)识别为时间要素。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。

准确率(P):

$$P = \frac{E_r}{E_t} \tag{3}$$

召回率(R):

$$R = \frac{E_r}{(E_a + E_t)} \tag{4}$$

F_1 值(F_1):

$$F_1 = \frac{2 * P * R}{P + R} \tag{5}$$

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。

表 5 CEC 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	5 583	7 523	4 937
Time	1 320	1 935	1 329
Location	1 035	1 607	1 476
Participant	1 195	1 802	2 928
ReportTime	314	332	332

表 6 CEC 要素自动标注实验结果

要素	准确率/%	召回率/%	F ₁ 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

5.2 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 5 954 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 7 523 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。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,在后期需要进一步改进。

表 7 CEC 事件识别对比

人工标注事件个数	5 954
自动标注事件个数	7 523

5.3 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验证,采用 Stanford Named Entity Recognizer (NER)^[15] 进行对比实验。Stanford NER 也叫条件随机场分类器,是一个 Java 实现的命名实体识别程序(以下简称 NER)。NER 基于一个训练而得的 Model 工作,用于训练的数据即大量人工标记好的文本,理论上用于训练的数据量越大,NER 的识别效果就越好。但是对于中文识别,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 语料进行识别,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,实验结果如表 8、表 9 所示。

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	1 802	88.40
PERSON∩Participant	246			
MISC∩Participant	905			
LOC∩Location	128	Location	1 607	89.61
GPE∩Location	1 110			
MISC∩Location	202			

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	1 802	88.51
PERSON∩Participant	238			
MISC∩Participant	917			
LOC∩Location	137	Location	1 607	89.61
GPE∩Location	1 111			
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。

5.4 实验 4—CEEC 语料库要素识别

CEEC^① (Chinese Environment Emergency Corpus)是利用人工标注所构建的环境污染类突发事件语料库,共包括六类,分别是: 噪声污染、土壤污染、水污染、海洋污染、空气污染和社会效应,合计 100 篇。参考实验 1 的步骤,经过对 CEEC 的自动标注实验,标注要素个数统计如表 10 所示,实验结果如表 11 所示。

表 10 CEEC 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	627	936	1 134
Time	241	344	257
Location	212	336	310
Participant	191	288	464
ReportTime	100	100	100

表 11 CEEC 要素自动标注实验结果

要素	准确率/%	召回率/%	F ₁ 值/%
Denoter	66.99	60.58	63.62
Time	70.06	80.20	74.79
Location	63.10	65.64	64.34
Participant	66.32	50.80	57.53
ReportTime	100.00	100.00	100.00

由实验结果可以看出,ReportTime 和 Time 要素格式统一,成分单一,识别效果较理想;由于事件触发词表是基于 CEC 所构建,将其用于识别 CEEC 类语料,亦具有较高的识别率;Participant 通常包括施动者(人)、参与者(机构、组织等),成分复杂,因此自动识别率偏低。

6 结束语

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,可以极大地提高标注速度。

(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。

本文方法仍存在需改进的地方,主要体现在触发词和事件要素的自动识别准确度尚未达到非常理想的程度,另外事件关系的识别及推理还需深入研究。

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價值[J].传媒,2014(2):64-66.

[2] LI Xiang, LIU Gang, LING Anhong, 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 Qingquan, RONG Lili. 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].上海大学博士学位论文,2010.

[5] 赵军,刘康,周光有,等.开放式文本信息抽取[J].中文信息学报,2011,25(6):98-110.

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

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

① <https://github.com/daselab/CEEC-Corpus>

[8] 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>.

[9] 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.

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


[11] 刘茂福, 李妍, 姬东鸿. 基于事件语义特征的中文文本蕴含识别[J]. 中文信息学报, 2013, 27(5): 129-136.

[12] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform [C]//Proceedings of the Coling 2010: Demonstrations. 2010. 08, pp13-16, Beijing, China


[13] 同义词词林扩展版 [A Thesaurus of Chinese Words][DB], http://www.ltp-cloud.com/download/#down_cilin.

[14] 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.

[15] Jenny Rose Finkel, Trond Grenager, Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling[C]//Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, 2005: 363-370.



刘炜(1978—), 博士, 副研究员, 主要研究领域为知识表示与推理, 语义网与本体技术。
E-mail: liuw@shu.edu.cn



王旭(1989—), 硕士研究生, 主要研究领域为自然语言处理与机器学习。
E-mail: wangx89@126.com



张雨嘉(1992—), 硕士研究生, 主要研究领域为自然语言处理, 知识表示, 机器学习, 统计机器翻译等。
E-mail: yujia_zhang@shu.edu.cn



(上接第 75 页)

[12] Yachao Li, Hongzhi Yu. Study on Tibetan Word Segmentation as Syllable Tagging [C]//Proceedings of Natural Language Processing and Chinese Computing (NLP&CC 2013). 2013, 11: 363-369.

[13] Haodi Feng, Kang Chen, Xiaotie Deng, et al. Accessor variety criteria for Chinese word extraction. Computational Linguistics [J]. 2004, 30(1): 75-93.

[14] Paul Cohen, Brent Heeringa, Niall Adams. An unsupervised algorithm for segmenting categorical time-series into episodes [C]//Proceedings of Pattern Detection and Discovery. 2002: 117-133.

[15] Paul Cohen, Brent Heeringa, Niall Adams. An unsupervised algorithm for segmenting categorical time-series into episodes [J]. Pattern Detection and Discovery. 2002: 117-133.

[16] Kumiko Tanaka-Ishii, Zhihui Jin. From phoneme to morpheme: Another verification using a corpus [C]//Proceedings of the 21st International Conference on Computer Processing of Oriental Languages. 2011: 234-244.



李亚超(1986—), 讲师, 主要研究领域为机器翻译、词法分析、少数民族语言文字信息处理。
E-mail: liyc7711@gmail.com



加羊吉(1985—), 博士, 副教授, 主要研究领域为藏文信息处理。
E-mail: 236164976@qq.com



江静(1988—), 助教, 主要研究领域为复杂网络。
E-mail: 506775848@qq.com