文章编号：1003-0077(2016)04-0167-09

一种基于事件本体的文本事件要素提取方法

刘 炜，刘菲京，王 东，刘宗田

（上海大学 计算机科学与工程学院，上海 200444）

摘 要：在事件信息的抽取中，事件要素的提取是一个难点。现有的事件要素抽取主要是基于机器学习的方法，这类方法容易受到语料稀疏性的影响。该文提出一种基于事件本体的事件要素提取方法，该方法将事件要素推理分为两步：一、通过事件要素词和事件指示词的位置关系来初步填充要素值，并将得出的置信度较高的事件作为种子事件；二、利用第一步得出的种子事件，查询事件本体中的事件类约束和基于事件非分类关系的推理规则，并对要素进行推理，进一步对事件要素进行填充和修正。实验结果表明，该方法能较好地提升事件要素提取的准确度。

关键词：事件本体；事件要素；事件要素推理

中图分类号：TP391 **文献标识码**：A

A Text Event Elements Extraction Method Based on Event Ontology

LIU Wei, LIU Feijing, WANG Dong, LIU Zongtian

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

Abstract: Extraction of event elements is a challenge in event-based information extraction. Currently, the main solutions are based on machine learning method which is subject to the corpus sparsity. This paper proposes an event element extraction method based on event ontology. Event elements reasoning process includes two steps: Firstly, elements values are initially complemented according to positional relations between event elements words and event indicators words, selecting the event with the highest confidence as the seed event; Secondly, search the seed events to for their event classes restrictions and non-taxonomic relations from event ontology, to complement and revise event elements. The experimental results show that this method can improve the accuracy of event elements extraction.

Key words: Event Ontology; Event Elements; Event Elements Reasoning

1 引言

在自然语言处理领域，“事件”可以描述比“概念”粒度更大的、动态的、具有完整意义的结构化知识，更加符合人类的认知规律，是近年来倍受关注的一种知识模型。因此，从自然语言中抽取事件信息也显得越来越重要。事件信息抽取中的关注点包括两个方面，即发生了什么事情（事件识别）和与事件密切相关的信息（事件要素信息，如时间、地点和人物）。同时，借助基于事件的文本表示方法，通过事件要素建立事件之间的关系，把描述这些事件的文档联系起来，可实现如文本分类、话题检测与跟踪等任务。

目前，事件要素的识别和抽取主要采用机器学习的方法，如文献[1-3]中的方法。这种方法将事件抽取任务转化为分类问题，虽然具有较好的鲁棒性，但分类器的构建、特征的发现和选择，以及作为模型训练基础的大规模语料库的标注工作都需要大量的人力和时间花费。针对机器学习方法的不足，本文提出一种基于事件本体的文本事件要素提取方法，该方法使机器能够模仿人的阅读习惯，通过事件本体对事件信息进行联想，对地点、时间、主体、客体四个事件要素进行推理。

事件的抽取分为事件类型的识别和事件要素抽取等任务。事件类型识别的目的是将事件分类，而

收稿日期：2014-10-15 定稿日期：2015-05-20

基金项目：国家自然科学基金(61305053)；国家自然科学基金(61273328)

要素的识别是为了事件信息的补全，将事件的发生时间、地点、人物等信息填充到相应事件中。现有的事件抽取方法中，利用最大熵分类器对事件的命名实体、时间等要素进行识别[4-5]是较常见的做法。文献[6]结合 MegaM 和 TiMBL 两种机器学习方法在 ACE 语料上均取得了不错的效果，但较小语料规模造成了一定的数据稀疏。文献[7]通过对事件类别的确定获得了该类事件的模板，将事件要素识别转化为二元分类问题，从一定程度上提高了事件要素识别效果，但还是不可避免的受到语料规模限制。文献[8]采用基于关键词与触发词相结合的过滤方法进行事件类型的识别，进而采用基于最大熵分类方法对事件元素进行识别，但该方法对学习语料的依赖性较强。此外，模式匹配的方法在事件要素识别中也被经常使用。其思路是建立一系列的模式，把句子与模板进行匹配达到事件识别与抽取的目的。这种方法只适合于特定的领域，缺乏通用性。典型例子是针对开放域的事件抽取系统 FSA[9]。文献[10-11]采用基于规则的方法分别从金融领域和突发事件领域抽取事件要素。规则的制定需要人工参与，不同规则之间还有可能出现冲突。文献[12]采用多层模式匹配的方法在 ACE 中文语料上识别事件要素，但所采用的规则有限导致识别效果不够理想。在准确率上，模式匹配的方法一般比机器学习的方法高，但过于依赖具体领域，可移植性差。

基于现有方法中存在的问题，本文采用基于事件本体的要素推理方法来实现事件要素的提取，首先根据词语位置关系初次填充要素，然后借助事件本体通过少量的推理规则来进行事件要素推理和填充。此方法可以从一定程度上解决对语料的依赖和规则制定的问题。

2 事件本体的构建

本文以文献[13]所提出的事件及事件关系概念为基础，并在此基础上提出上层事件本体结构，由此来构建针对事件要素提取的事件本体。以下对文献[13]中所提出的事件、事件类和事件关系等概念进行简单介绍。

2.1 事件相关定义

定义 1 事件(Event)和事件类，事件是指在某个特定的时间和环境下发生的，由若干角色参与，表现出若干动作特征的一件事情。事件类(Event Class)指具有共同特征的事件的集合。事件在形式上定义为一个六元组结构：

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

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。本文主要对事件的对象(主体和客体)、时间以及地点要素进行推理。

定义 2 事件关系，指的是存在于事件或事件类之间的分类关系和非分类关系。事件分类关系即事件类的包含关系，例如，自然灾害类包含地震类。事件非分类关系指的是事件或事件类之间存在的因果关系、跟随关系、并发关系和组成关系。通常这些关系既存在于事件实例之间，也存在于事件类之间。关于事件关系的语义定义见文献[13]。

2.2 上层事件本体结构

为支持事件要素的推理，在文献[13]事件本体结构的基础上，构建一个上层事件本体结构。上层事件本体结构定义了事件的分类层次结构，如表 1 所示。

表 1 上层事件本体分类结构

1 *Class*：*HumanEvent*
1.1*Class*：*SinglePersonEvent*
1.1.1*Class*：*PersonObject_SinglePersonEvent*
1.1.1.1 *Class*：*Continue_PO_SinglePersonEvent*
1.1.1.2*Class*：*Instant_PO_SinglePersonEvent*
1.1.2 *Class*：*NonObject_SinglePersonEvent*
1.1.2.1 *Class*：*Continue_NO_SinglePersonEvent*
1.1.2.2 *Class*：*Instant_NO_SinglePersonEvent*
1.1.3*Class*：*NonPersonObject_SinglePersonEvent*
1.1.3.1 *Class*：*Continue_NPO_SinglePersonEvent*
1.1.3.2*Class*：*Instant_NPO_SinglePersonEvent*
1.2 *Class*：*PublicEvent*
1.2.1 *Class*：*NonPersonObject_PublicEvent*
1.2.1.1 *Class*：*Instant_NPO_PublicEvent*
1.2.1.2 *Class*：*Continue_NPO_PublicEvent*
1.2.2*Class*：*PersonObject_PublicEvent*
1.2.2.1*Class*：*Continue_PO_PublicEvent*
1.2.2.2*Class*：*Instant_PO_PublicEvent*
1.2.3*Class*：*NonObject_PublicEvent*
1.2.3.1 *Class*：*Continue_NO_PublicEvent*
1.2.3.2 *Class*：*Instant_NO_PublicEvent*
2 *Class*：*NatureEvent*
2.1*Class*：*NonNatureForceEvent*
2.1.1 *Class*：*PersonObject__NonNatureForceEvent*

续表

2.1.1.1 *Class*：*Continue_PO_NonNatureForceEvent*
2.1.1.2 *Class*：*Instant_PO_NonNatureForceEvent*
2.1.2 *Class*：*NonPersonObject _NonNatureForceEvent*
2.1.2.1 *Class*：*Continue_NPO_NonNatureForceEvent*
2.1.2.2 *Class*：*Instant_NPO_NonNatureForceEvent*
2.1.3 *Class*：*NonObject_NonNatureForceEvent*
2.1.3.1 *Class*：*Continue_NO_NonNatureForceEvent*
2.1.3.2 *Class*：*Instant_NO_NonNatureForceEvent*
2.2 *Class*：*NatureForceEvent*
2.2.1 *Class*：*PersonObject__NatureForceEvent*
2.2.1.1 *Class*：*Continue_NatureForceEvent*
2.2.1.2 *Class*：*Instant_NatureForceEvent*
2.2.2 *Class*：*NonPersonObject __NatureForceEvent*
2.2.2.1 *Class*：*Continue_NPO_NatureForceEvent*
2.2.2.2 *Class*：*Instant_NPO_NatureForceEvent*
2.2.3 *Class*：*NonObject__NatureForceEvent*
2.2.3.1 *Class*：*Continue_NO_NatureForceEvent*
2.2.3.2 *Class*：*Instant_NO_NatureForceEvent*

上层事件结构的第一层根据事件类的主体类别划分为两大类：人类事件类和自然事件类。

第二层进一步地根据事件类的主体数量把人类事件类划分为个人事件类和公共事件类。多人参与的事件类为公共事件类，而单个人参与的事件类为个人事件。例如，驾驶和交通事故的区别。自然事件类中的第二层分为自然力事件和非自然力事件，自然力事件的主体通常是大自然，如台风、山洪暴发等；非自然力事件的主体是一切除了人类和大自然的物体，可以是大自然中的物质，如一氧化碳、石头等，也可以是人类社会生产出来的物品，如高速公路、汽车等。

在上层本体的第三层，人类事件类根据事件类的客体划分为人类客体事件类、非人类客体事件类和不及物事件类。不及物事件类一般描述事件主体内部状态的变化，不会对其他事物产生影响，如生病和死亡等。自然事件类的第三层也是根据事件类的客体进行划分，自然力事件类下面分为人类客体自然力事件类、非人类客体自然力事件类以及不及物自然力事件类，非自然力事件类也是同样的划分方法。但是实际情况下，自然事件的客体往往是可以忽略的，因为这些事件大多数是自发事件，例如，地震事件和汽车爆炸事件。

第四层则是在第三层事件类基础上根据时间来划分。根据事件的时间要素，可以分为瞬时事件和持续事件。这样划分有利于分析事件的包含和组成关系，因为如果持续事件的时间较长，则在该持续事件发生的时间段内可能包含了瞬时事件和其他持续事件。

2.3　事件本体的建立

在事件本体的开发过程中，上层事件本体是不需要建立的，都是被预先定义好的抽象类。新建的具体事件类则需要根据事件类要素来进行划分，使之归类到上层事件本体中的某个事件类。并建立具体事件类之间的关系，形成具体的事件本体。

具体事件类通过扩展 OWL 语言进行描述。事件关系包括并发（*concur*）、因果（*cause*）、跟随（*follow*）和组成（*is_part_of*）几种关系。这些事件关系在 OWL 中通过 *ObjectProperty* 类型建立，每个事件类都有若干个 *ObjectProperty* 类型的属性，如因果、跟随等，用 *restriction* 来限制一个事件类在某个 *ObjectProperty* 类型上与其他事件类的一一对应关系。*Restriction* 定义了三种类型：*all values from*、*some values from*、*has value*。*All values from* 表示指定属性的所有可能取值都只能从指定的类中选取。*Some values from* 表示指定属性的部分值从指定的类中选取，而 *has value* 表示必须取规定的特定值。例如，倒塌事件类定义了一个表示因果关系的 *ObjectProperty*，*cause* 属性的约束 *restriction* 为“*some values from* 地震”，即表示“*倒塌*”事件部分是由于“*地震*”引起的。这种方式不仅能够描述事件类的关系类型，还能够描述事件关系的概率。图 1 是包含了上层事件结构的事件本体模型。

3　事件要素的推理和识别流程

事件本体是一个包含所有的事件类及事件类之间关系的集合。特定事件的要素约束条件可以通过查询事件本体得到。但是只是通过要素的约束条件很难在符合条件的大量要素中完成要素识别任务。对于事件要素的识别，可以根据上下文中与某个事件相关联的要素来推理出这个事件的相关信息。本文模拟这种联系上下文的方式来制定推理规则，使用事件关系来建立文章中事件的联系。本节分析了事件类之间的关系及其各自对要素推理的作用，分别定义了推理规则，并描述了要素识别的流程。

3.1　针对四类关系的要素推理规则

本文的中事件类之间的关系分为分类关系和非

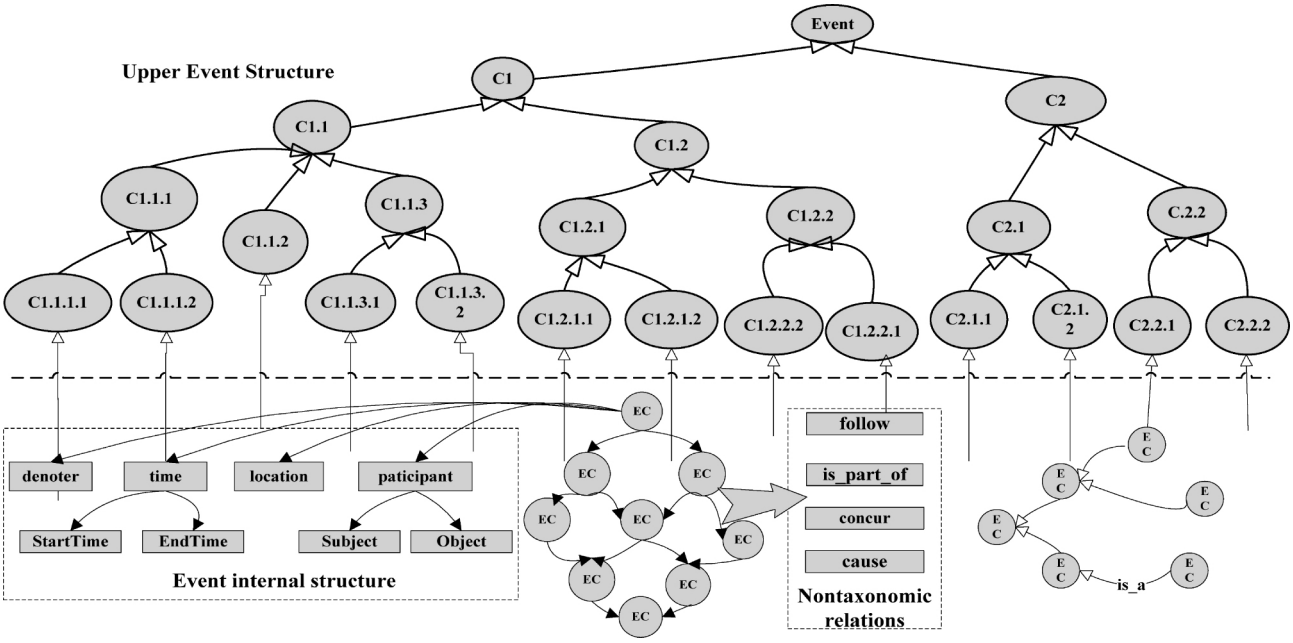

图1 事件本体结构图

分类关系，两类关系对要素推理的作用各不相同。对于分类关系，根据查询到某事件在上层事件结构中所属的抽象事件类，可以获得该事件的要素约束条件。例如，一个事件类（如打雷）属于瞬时自然力事件类，那它的开始时间和结束时间相同，而且它的客体为空。

非分类关系在文本事件要素的推理过程中起到联接上下文的作用，是要素推理的主要内容。经过对上层本体中所有第四层次的事件类型特征的研究以及大量案列的分析，根据事件之间的关系，我们针对每一种事件类型组合，分别提出了一组事件要素的推理规则，包括对地点、时间、主体和客体四个要素的推理，形成一个事件要素推理规则库。表2是针对两个 *Continue_PO_PublicEvent* 事件类型（简称为 *CPOPE* 类型，即存在关系的两个事件都是属于多人参与的公共持续事件）事件之间的关系所制定的12条推理规则，同样，我们针对其它的不同类型事件之间的关系组合也可以分别制定推理规则。在这些推理规则中，$Sub(e_i)$ 表示 e_i 的主体对象，$Obj(e_i)$ 表示客体对象，$P(e_i)$ 表示事件 e_i 的地点要素，$ST(e_i)$ 表示事件开始时间，$ET(e_i)$ 表示结束时间。表2中的推理规则解释如下。

（1）组成关系

存在组成关系的两个事件通常具有相同的地点要素和主体要素，如“*救助*”和“*现场施救*”的主体都是“*医疗人员*”。在组成关系中，小事件的客体通常是大事件客体的一部分，例如，“*现场施救*”的客体“*伤员*”是“*救助*”的客体“*所有在事故现场受伤的人*”的组成部分。由以上规则还可以推出组成事件类的兄弟子事件类通常具有某些相同的要素（如主体和地点），例如，“*救助*”的子事件“*现场施救*”和“*赶赴现场*”具有相同的主体“*医疗人员*”和相同的地点“*事故现场*”。总结归纳可得到表3中的规则 a 到规则 d，即对于 *CPOPE* 类型的事件 e_1 和事件 e_2，若 e_1 是 e_2 的组成事件，规则 a 表示 e_2 的时间区间包含 e_1 的时间区间；规则 b 表示事件 e_1 和 e_2 在相同的地点发生；规则 c 表示事件 e_1 和 e_2 具有相同的主体；规则 d 表示 e_1 的客体是 e_2 客体的一部分。

（2）因果关系

对于存在因果关系的两个 *CPOPE* 类型事件，其发生的地点往往是相同的，时间上起因事件通常发生在结果事件之前。规则 e 表示起因事件 e_1 的起始时间在结果事件 e_2 的起始时间之前；f 表示起因事件 e_1 和结果事件 e_2 通常发生在相同地点；g 表示起因事件 e_1 的客体通常是结果事件 e_2 的主体。

（3）跟随关系

对于存在跟随关系的两个 *CPOPE* 类型事件，其发生的时间通常有先后，而且相隔时间较短，两个事件在时间区间上不存在重叠。此外，两个事件类一般具有相同的主体和地点要素。若 e_2 跟随 e_1 发生，规则 h 表示事件 e_1 结束之后事件 e_2 才发生；i 表示事件 e_1 和事件 e_2 的发生地点是相同的；j 表示事

件 e_1 和事件 e_2 的主体是相同的。

（4）并发关系

存在并发关系的两个 $CPOPE$ 类型事件通常是同时发生，两个事件的时间要素和地点要素通常是相同的。规则 k 表示存在并发关系的两个 $CPOPE$ 类型事件的地点要素相同。规则 l 表示存在并发关系的两个 $CPOPE$ 类型事件的发生时间存在重叠。

表 2　针对 *CPOPE*×*CPOPE* 事件关系的要素推理规则

a.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ is\_part\_of\ EC_2) \Rightarrow (ST(e_1) \geqslant ST(e_2)) \cap (ET(e_1) \leqslant ET(e_2))$
b.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ is\_part\_of\ EC_2) \Rightarrow P(e_1) = P(e_2)$
c.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ is\_part\_of\ EC_2) \Rightarrow Sub(e_1) = Sub(e_2)$
d.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ is\_part\_of\ EC_2) \Rightarrow Obj(e_1) \subseteq Obj(e_2)$
e.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ cause\ EC_2) \Rightarrow ST(e_1) < ST(e_2)$
f.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ cause\ EC_2) \Rightarrow P(e_1) = P(e_2)$
g.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ cause\ EC_2) \Rightarrow Obj(e_1) = Sub(e_2)$
h.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_2\ follow\ EC_1) \Rightarrow ET(e_1) < ST(e_2)$
i.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_2\ follow\ EC_1) \Rightarrow P(e_1) = P(e_2)$
j.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_2\ follow\ EC_1) \Rightarrow Sub(e_1) = Sub(e_2)$
k.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ concur\ EC_2) \Rightarrow P(e_1) = P(e_2)$
l.	$(e_1 \in EC_1) \cap (e_2 \in EC_2) \cap (EC_1\ concur\ EC_2) \Rightarrow (ST(e_1) \leqslant ST(e_2)) \cap (ET(e_1) \geqslant ET(e_2))$

3.2　事件要素识别过程

本文主要针对新闻报道文本中四个要素（地点、时间、主体、客体）进行识别和填充。对于一篇文章，抽取出其中所有命名实体，地点词、人物词和时间词等能够表示事件要素的词语，可构建一个二维矩阵，纵向维度的各行表示不同事件，横向维度的各列表示事件要素词。矩阵中的各个数值代表不同的要素类型表征：0 表示要素不隶属于该事件，1 表示地点要素，2 和 3 分别表示开始时间和结束时间要素，4 和 5 分别表示主体和客体对象要素。通过不断更新这个矩阵，实现事件要素的填充。例如，A_{ij} 描述了一篇文章中所有事件所构成的矩阵。

$$Aij = \begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \end{array} \begin{array}{c} \begin{array}{cccccccccc} w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 & w_{10} \end{array} \\ \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 3 & 4 & 5 & 0 & 0 & 2 & 0 & 0 \\ 0 & 2 & 3 & 0 & 0 & 0 & 1 & 0 & 4 & 5 \\ 0 & 0 & 3 & 4 & 5 & 2 & 1 & 0 & 0 & 0 \\ 1 & 0 & 3 & 4 & 5 & 2 & 0 & 0 & 0 & 0 \\ 0 & 2 & 3 & 5 & 0 & 0 & 1 & 0 & 4 & 0 \\ 0 & 2 & 3 & 5 & 0 & 0 & 1 & 0 & 4 & 0 \\ 1 & 2 & 3 & 4 & 0 & 0 & 0 & 0 & 0 & 5 \end{bmatrix} \end{array}$$

事件要素识别过程主要包含三个阶段：数据的预处理、基于词位置的要素初步填充、要素的推理。

数据的预处理首先要对文章进行分词并手工修正分词过细的结果，然后标出事件触发词和对应的事件要素词。以句子为单位标出词在句子中的序号，以便能够在后面步骤中计算词的位置关系。

初步填充阶段需要在预处理阶段标出了词语在文中的段落序号、句子序号、和词语序号的基础上，计算触发词和要素词的距离关系，将距离最近的词作为要素初步填充的结果。对于文中的事件 e，初步填充要素的步骤见表 3。这里的 α、β 和 γ 是用于计算置信度的权值，分别代表要素词和事件触发词在同一个句子、同一个段落、不在一个段落。为了保证置信度随着距离的增加而减小（一般来说，要素词和触发词在相同句子的置信度比在不同句子的置信度大），将三个置信度权值分别取值为 $\alpha = 100$，$\beta = 10$，$\gamma = 1$。

表 3　要素初步填充步骤

步骤 1	从要素词的链表中取出要素词 $element1$，利用函数 $confident(e,i)$ 计算该事件的第 i 个要素的置信度，初始值为 0；跳转到步骤 2；
步骤 2	计算 $element1$ 的置信度： *if*（e 和 $element1$ 位于相同的句子中） 　$conf = \alpha / |e$ 的词号 − $element1$ 的词号|； *else if*（e 和 $element1$ 在相同的段落中） 　$conf = \beta / |e$ 的词号 − $element1$ 的词号|； *else* 　$conf = \gamma / |e$ 的词号 − $element1$ 的词号|； 转到步骤 3；
步骤 3	将算出的置信度与之前的进行比较，如果置信度提高了，则更新要素值，并且更新 $confident(e,i) = conf$； 否则，保持原来的置信度和要素值不变，转步骤 4；
步骤 4	跳转到步骤 1，继续取出事件要素，直到所有的事件要素遍历完。

第三阶段利用第二阶段填充的结果，对事件要

素进行推理。首先查询事件所属的上层事件类，得到该事件的要素约束条件，例如，有些事件的某个要素是缺省的则不填充，有些事件的主体只能是人，则选择表示人物的命名实体来填充。为了保证推理的准确性，需要从初步填充结果中选择置信度最大的事件作为种子事件进行推理。把种子事件作为输入对事件本体进行查询，找到与其存在非分类关系的事件类，然后查询每一对关联的两个事件类的上层事件类型，根据所关联的两个事件类的上层类型定位要素推理规则库，接下来利用推理规则进行推理。

3.3 实例分析

以下是一段半自动标注后的新闻，e_i表示事件触发词，l_i表示地点词，t_i表示时间词，p_i表示参与者（包括了主体和客体）：

新快报讯，8月20日早上6点(t_1)，阿尔及利亚以东150公里的卜伊拉(l_1)发生汽车炸弹(p_1)爆炸(e_1)事件，造成11人(p_2)死亡(e_2)。

当地媒体报道称，包括4名军事人员在内的31人(p_3)受伤(e_3)。目前(t_2)，当地(l_2)正对伤者(p_4)进行救治(e_4)。

第一步，通过词语位置远近填充的事件要素如表4，$conf$为置信度。

表4 通过词语位置远近填充的事件要素

事件	LOC	ST	ET	SUBJECT	OBJECT	conf
e_1	l_1	t_1	t_1	p_1	p_1	291.7
e_2	l_1	t_1	t_1	p_2	p_2	247.6
e_3	l_1	t_2	t_2	p_3	p_3	221
e_4	l_2	t_2	t_2	p_4	p_4	165

第二步，通过事件本体中的事件类约束进行推理。

e_1 is_a $Instant\_NonNatureForceEvent$

$\Rightarrow e_1.ST=e_1.ET=t_1, e_1.OBJECT=null$ (1)

式(1)说明，把e_1映射到本体中得到其上层的事件类型为$Instant_NonNatureForceEvent$，一是可以得出$e_1$是瞬时事件，则开始时间和结束时间相同，二是该事件描述的是主体自身的变化，没有客体。由于第一步会把距离最近的要素词p_1填充为e_1的客体，不符合$Instant_NonNatureForceEvent$类型没有客体的约束，所以第二步修正$e_1$的客体为空。同理，可以得出：

e_2 is_a $Instant\_NonObject\_SinglePersonEvent$

$\Rightarrow e_2.ST=e_1.ET=t_1, e_2.OBJECT=null$

e_3 is_a $Continue\_NonObject\_SinglePersonEvent$

$\Rightarrow e_3.OBJECT=null, e_3.ET>e_3.ST=t_1$

e_4 is_a $Continue\_PersonObject\_PublicEvent$

$\Rightarrow e_4.OBJECT=$ 伤者，$e_4.ET>e_4.ST=t_2$

其中，$e_4.SUBJECT=$*医疗人员*，根据具体的“*救治*”事件类得出它在本体中的主体要素。

第三步，从事件本体中查询获得以下几种事件关系：$e_1 cause\ e_2$、$e_1 cause\ e_3$、$e_2 concur\ e_3$、$e_3 cause\ e_4$。将e_1作为种子事件，根据事件1和事件2的类型，在推理规则库中查找相应的规则，然后对其他事件的事件要素进行推理：

$e_1 cause\ e_2 \Rightarrow e_1.ST< e_2.ST \Rightarrow e_2.ST = e_2.ET = t_1+$

$e_2\ concur\ e_3 \Rightarrow e_2.ST=e_3.ST \Rightarrow e_3.ST = e_3.ET = t_1+$

$e_3\ cause\ e_4 \Rightarrow e_3.ST< e_4.ST \Rightarrow e_4.ST = t_1++ (t_2=t_1++)$

$e_3\ cause\ e_4 \Rightarrow e_3.LOC= e_4.LOC (l_2=l_1) \Rightarrow e_4.LOC = l_1$

最后得到的结果如表5所示。

表5 要素推理结果

事件	LOC	ST	ET	SUBJECT	OBJECT
e_1	l_1	t_1	t_1	p_1	*null*
e_2	l_1	t_1+	t_1	p_2	*null*
e_3	l_1	t_1+	t_1++	p_3	*null*
e_4	l_1	t_1++	t_1+++	*医务人员*	P_4

可以看出，通过推理把本身没有对象要素的事件的对象值设置为空。更新了事件发生的时间，并且从本体中填充了默认的要素“*医务人员*”，将如“*目前*”、“*当地*”等相对时间和地点推理出其绝对的事件和地点，在一定程度上填充了事件要素。

4 实验和分析

4.1 数据集和评价标准

本实验的数据采用突发事件语料库(Chinese Emergency Corpus, CEC)[14]，其中包含了地震、火灾、交通事故、恐怖袭击以及食物中毒五类突发事件

的语料共 332 篇。事件本体采用文献[15]中构建的突发事件本体，包含事件类 421 个、事件间的因果和跟随等关系 307 个。

通过准确率(precision)、召回率(recall)、F 值(F-Measure)这三个标准来评价要素填充的效果。

其中，准确率是计算正确填充要素的事件数占所有事件总数的比例。

$$precision = \frac{\text{正确填充要素的事件数}}{\text{事件的总数}}$$

召回率用来计算正确填充某要素的事件数占包含该要素的事件总数的比例。

$$recall = \frac{\text{正确填充某要素的事件数}}{\text{包含该要素的事件总数}}$$

F1 值的计算方法如式(1)所示。

$$F1 = \frac{2 \times recall \times precision}{recall\,precision} \tag{1}$$

4.2　实验设计

实验选取 CEC 语料中的若干个事件，分别进行主体、客体、地点和时间要素的填充，实验设计为两个部分。

实验一：使用邻近的要素进行填充，也就是根据事件触发词和要素词之间的位置关系来填充。对于报道中的一些格式化的词语，比如“据新华社报道”、“某人说”，如果将其作为要素补充的候选，会对实验结果产生干扰。此外，这一类事件的描写通常不是用来描述事情的发展情况。所以，这类事件在实验中会被剔除掉。利用分词工具标注的人称代词和命名实体等概念，以及地名和时间等要素也会因为表示的格式不同带来判断不一致情况，要对这些词语进行统一。有些事件的客体是缺省的，例如，“海啸”的客体为空，所以这一类事件的客体不需要统计结果。

实验二：根据本文所提出的事件关系推理规则，利用推理的结果来填充事件要素。要素推理需要选取一个置信度较高的事件通过非分类关系推理出关联事件的要素。文章的标题和第一段中提到的事件通常不会把事件要素缺省，对于这些事件，从第一部分实验中得到数据较为准确。所以实验二将在实验一得到较优结果的基础上分两种情况进行实验：(1)选择置信度最高且出现在第一段中的事件作为种子事件进行推理；(2)选择置信度最高且出现在其他段落中的事件作为种子事件进行推理。

4.3　实验结果分析

实验一对 CEC 语料中随机抽取的 195 个事件的统计结果如表 6 所示。其中，对于地点要素和时间要素的填充结果，准确率、召回率和 F1 值都超过了 60%。由此可见，利用触发词和事件要素词的位置关系实现对这两种要素的抽取，能初步达到理想的效果。文本中地点要素词和事件要素词所涉及的范围可以根据其在篇章结构的远近来初步判断。而对主体和客体的填充效果不如时间、地点要素，原因包括三点：第一，文章中主体和客体词出现的次数明显多于主体和客体，容易造成其在句子中的分布互相干扰；第二，主体词可能在其他事件中充当客体，客体词也可能在一些事件中成为主体，即主体和客体的标注不像地点词和时间词那样明确；第三，一些事件的主体和客体会出现多个，但是只能填充其中的一个。

表 6　不同关系的邻近要素填充结果

要素	Precision/%	Recall/%	F1 measure/%
地点	65.8	62.5	64.1
时间	60.3	66.6	63.3
主体	48.5	72.3	58.1
客体	54.3	52.7	53.5

从实验一中得出的置信度最高的事件分布，如图 2。

根据图 2 可以看出，置信度最高和次高的事件通常会出现在文章的第一自然段，其次是第二自然段，其他的自然段分布比较均匀。通常一篇文章的核心事件都是分布在第一自然段，叙述也较为详细，而且会在一句话中交待该事件的主要要素。分布在其他自然段的置信度高的事件则通常不是文章的核心事件。

实验一只运用了词语间的位置关系，事件的类型、上下文关系等因素没有考虑进去，所以实验二利用本体查询事件关系和上层事件类型，通过基于非分类关系推理更新第一步的填充结果。实验结果如表 7。

表 7　针对不同种子事件的要素推理填充结果

要素	FirstPara_Event			OtherPara_Event		
	P/%	R/%	F1/%	P/%	R/%	F1/%
地点	78.3	82.3	80.3	73.5	80.7	76.9
时间	80.9	76.2	78.5	75.7	73.6	74.6
主体	69.3	73.7	71.4	68.0	66.9	67.4
客体	70.5	68.6	69.5	65.6	65.3	64.5

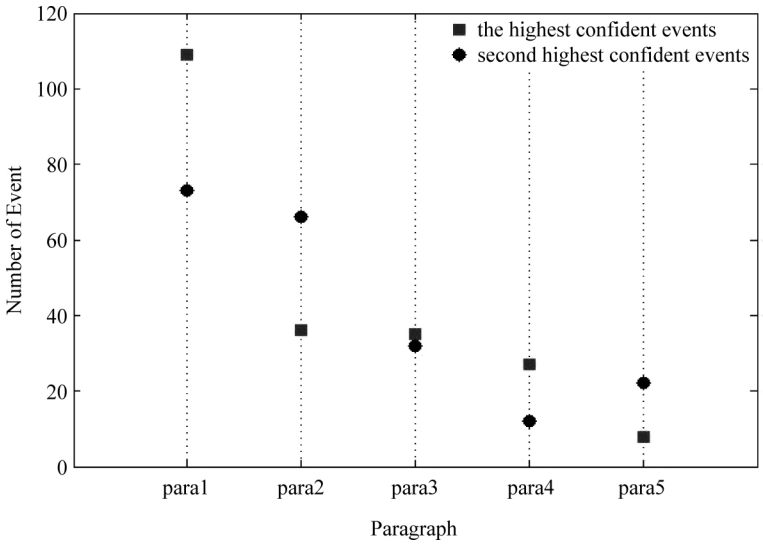

图2 置信度值最高和次高的事件数量分布图

由表7中可知，实验二的事件要素的识别效果比实验一显著提高。在各类要素的填充过程中，只要两个事件存在关系，即使在文中的位置关系并不相近，也能够推理出地点上的相似或者时间上的前后关系，识别效果得到提高，排除了词位置关系的干扰。主体和客体的识别上也避免了实验一中的不足，特别是当主体连续执行多个事件的时候，能够过滤掉句子中夹杂的客体词，把连续事件的主体统一填充为该主体。此外，实验二中事件映射到本体中的事件类，获得要素的约束条件，充分考虑了这些要素词语的类别和事件类要素缺省的情况。通过实验二从本体中获得的事件约束，可以对实验一的填充结果进行修正。对于非缺省的要素，通过事件类的要素约束选择出更合适的要素词能够提高实验的效果。在选取种子事件时，第一种情况的实验效果要略好于第二种情况。原因是第一种情况的种子事件往往是文中较为重要的事件，与其存在非分类关系的事件较多，所以通过这类种子事件进行修正的关联事件较多。从实验结果可以看出，利用基于事件非分类关系的推理能够有效地识别出事件要素，选择文章首段的事件作为种子，能够获得更好的实验效果。

与文献[2]采用基于机器学习的方法所获得的比较理想的实验结果对比，本文方法在地点和时间抽取的准确率和召回率略低，主体和客体的抽取准确率和召回率明显提高，综合四种不同要素的抽取，本文方法效果更理想，如表8所示。此外，由于本文采用的实验数据是具有普遍性的新闻文本，精确度和召回率相比文献[8]略低，但是降低了对特有语料的依赖性。

表8 准确率和召回率的比较

要素	本文方法		机器学习法	
	Precision/%	Recall/%	P/%	R/%
地点	78.3	82.3	78.2	88.7
时间	80.9	76.2	86.7	83.4
主体	69.3	73.7	59.1	32.7
客体	70.5	68.6	47.5	54.6

5 结论

本文针对传统事件要素识别方法所存在的缺点，提出了一种基于事件本体的文本事件要素识别和推理方法。建立了面向事件要素推理的包含两层结构的事件本体；定义了基于事件类关系的要素推理规则。相比基于规则的方法，本文方法所需要的规则数量更少；相比传统的基于机器学习的方法，本文的方法对语料的依赖性大大减弱，且对语料的数量没有具体的要求。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提高事件要素的识别效果。需要改进的地方体现在目前事件指示词和事件要素的自动识别准确度还不能达到较理想的程度，另外事件本体的结构影响要素识别效果，本体中事件要素的约束条件以及针对事件类关系的推理规则定义还需进一步完善。

参考文献

[1] Saeedi P, Faili H. Feature engineering using shallow parsing in argument classification of Persian verbs [C]//Proceedings of the 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012), 2012: 333-338.

[2] Wang W, Zhao D Y, Wang D. Chinese news event 5w1h elements extraction using semantic role labeling [C]//Proceedings of the Third International Symposium on Information Processing (ISIP), 2010: 484-489.

[3] **杨尔弘. 突发事件信息提取研究**[D]. **北京语言大学博士学位论文**, 2005.

[4] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text[C]//Proceedings of the 18th National Conference on Artificial Intelligence(AAAI 2002), 2002:786-791.

[5] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction[C]//Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, 2009: 209-212.

[6] Ahn D. The stages of event extraction[C]//Proceedings of COLING/ACL 2006 Workshop on Annotating and Reasoning about Time and Events, 2006: 1-8.

[7] **赵妍妍，秦兵，车万翔，等. 中文事件抽取技术研究**[J]. **中文信息学报**, 2008, 22(1): 3-8.

[8] **丁效，宋凡，秦兵，等. 音乐领域典型事件抽取方法研究**[J]. **中文信息学报**, 2011, 25(2): 15-20.

[9] Surdeanu M, Harabagiu S. Infrastructure for open-domain information extraction[C]//Proceedings of the Human Language Technology Conference (HLT 2002), 2002: 325-330.

[10] **周剑辉，苑春法，黄锦辉，等. 金融领域内信息抽取规则的自动获取**, in Advances in Computation of Oriental Languages[C]//Proceedings of the 20th International Conference on Computer Processing of Oriental Languages, Shenyang, China, 2003: 410-416.

[11] **梁晗，陈群秀，吴平博. 基于事件框架的信息抽取系统**[J]. **中文信息学报**, 2006, 20(2): 40-46.

[12] Tan H Y, Zhao T J, Zheng J H. Identification of Chinese event and their argument roles[C]//Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops, 2008: 14-19.

[13] **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.

[14] CEC-Corpus, https://github.com/daselab/CEC-Corpus[OL].

[15] **仲兆满. 事件本体及其在查询扩展中的应用**[D]. **上海大学博士学位论文**,2011.

刘炜(1978－)，**博士，副研究员，主要研究领域为语义本体、知识表示。**

E-mail: liuw@shu.edu.cn

王东(1986－)，**硕士研究生，主要研究领域为本体技术、事件知识表示等。**

E-mail: ming123@shu.edu.cn

刘菲京(1989－)，**硕士研究生，主要研究领域为事件本体建模及本体映射技术。**

E-mail: liufeijing0307@163.com