编者按: 中国中文信息学会于 2007 年 11 月在苏州大学成功地召开了“第四届全国信息检索与内容安全学术会议(NCIRCS-2007)”。会议的程序委员会向本刊推荐了 27 篇论文,并经作者仔细修改,编辑部得到授权,将在 2008 年第一、二期发表,以飨读者。

文章编号: 1003-0077(2008)01-0003-06

中文事件抽取技术研究

赵妍妍,秦 兵,车万翔,刘 挺

(哈尔滨工业大学 计算机学院 信息检索研究室,黑龙江 哈尔滨 150001)

摘 要: 事件抽取是信息抽取领域一个重要的研究方向,本文对事件抽取的两项关键技术——事件类别识别以及事件元素识别进行了深入研究。在事件类别识别阶段,本文采用了一种基于触发词扩展和二元分类相结合的方法;在事件元素识别阶段,本文采用了基于最大熵的多元分类的方法。这些方法很好的解决了事件抽取中训练实例正反例不平衡以及数据稀疏问题,取得了较好的系统性能。

关键词: 计算机应用;中文信息处理;事件抽取;事件类别识别;事件元素识别

中图分类号: TP391 **文献标识码**: A

Research on Chinese Event Extraction

ZHAO Yan-yan, QIN Bing, CHE Wan-xiang, LIU Ting

(Information Retrieval Laboratory, School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, Heilongjiang 150001, China)

Abstract: Event Extraction is an important research point in the area of Information Extraction. This paper makes an intensive study of the two stages of Chinese event extraction, namely event type recognition and event argument recognition. A novel method combining event trigger expansion and a binary classifier is presented in the step of event type recognition while in the step of argument recognition, one with multi-class classification based on maximum entropy is introduced. The above methods solved the data unbalanced problem in training model and the data sparseness problem brought by the small set of training data effectively, and finally our event extraction system achieved a better performance.

Key words: computer application; Chinese information processing; event extraction; event type recognition; event argument recognition

1 引言

事件抽取是信息抽取领域一个重要的研究方向。事件抽取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[1~3],自动问答[4],信息检索[4]等领域有着广泛的应用。

近些年来,事件抽取一直吸引着许多研究机构和研究者的注意力。MUC (Message Understanding Conference) 会议和 ACE (Automatic Content Extraction)会议是典型的含有事件抽取任务的评测会议。本文有关事件抽取的定义和实例来自于

收稿日期: 2007-05-31 **定稿日期**: 2007-12-03

基金项目: 国家自然科学基金资助项目(60575042, 60675034);国家 863 资助项目(2006AA01Z145)

作者简介: 赵妍妍(1983—),女,博士生,主要研究方向为信息抽取;秦兵(1968—),女,副教授,主要研究方向为信息抽取,多文档文摘;车万翔(1980—),男,讲师,主要研究方向为自然语言处理。

ACE[5]。根据定义,事件由事件触发词(Trigger)和描述事件结构的元素(Argument)构成。图 1 结合 ACE 的事件标注标准详细地表述了一个事件的构成。其中,“出生”是该事件的触发词,所触发的事件类别(Type)为 Life,子类别(Subtype)为 Be-Born。事件的三个组成元素“毛泽东”、“1893 年”、“湖南湘潭”,分别对应着该类(Life/ Be-Born)事件模板中的三个元素标签,即:Person、Time 以及 Place。

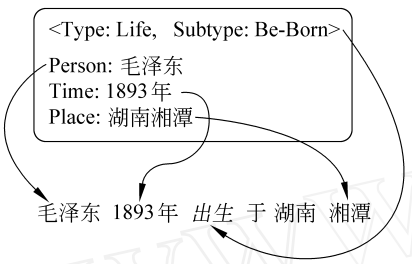

图 1 “出生”事件的基本组成要素

事件抽取任务可由下面两个主要步骤组成:

1. 事件类别识别:事件模板由事件的类别决定。ACE2005 定义了 8 种事件类别以及 33 种子类别,如表 1。每种事件类别/ 子类别(简称为“事件类别”)对应着唯一的事件模板,如表 2。

2. 事件元素识别:事件元素是指事件的参与者。根据所属的事件模板(如表 2),抽取相应的元素,并为其标上正确的元素标签。

表 1 ACE 定义的事件类别

Type	Subtype
Life	Born, Marry, Divorce, Injure, Die
Movement	Transport
Conflict	Attack, Demonstrate
Contact	Meet, Phone-Write
……	……

表 2 ACE 定义的事件模板

Type/ Subtype	Template
Life/ Be-Born	Person, Time-Within, Place
Business/ Merge-Org	Org, Time, Place
Contact/ Meet	Entity, Time, Duration, Place
……	……

2 相关工作及系统框架

事件抽取主要有两种方法:模式匹配的方法和机器学习的方法。模式匹配的方法是指对于某类事件的识别和抽取是在一些模式的指导下进行的,采用各种模式匹配算法将待抽取的句子和已经抽出的模板匹配[6,7]。例如 Surdeanu 和 Harabagiu 针对开放域的事件抽取系统——FSA[8]等。这种方法准确率较高,但往往依赖于具体领域,可移植性差。机

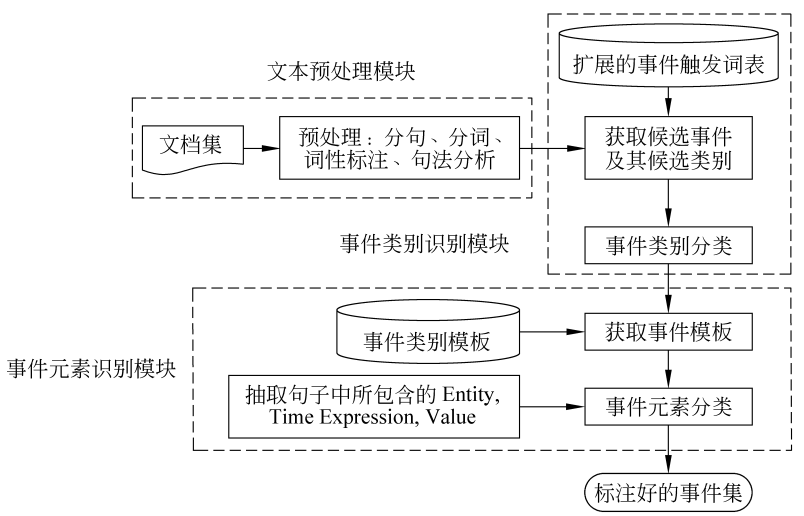

图 2 事件抽取系统框架图①

① 预处理采用了哈工大信息检索研究室开发的 LTP(Language Technology Platform)语言技术平台提供的技术模块。

器学习的方法把事件抽取任务看作分类问题,把主要的精力放在分类器的构建和特征的发现、选择上。相对而言,这种方法较为客观,不需要太多的人工干预和领域知识,因此目前的事件抽取研究多数采用机器学习的方法。Hai Leong Chieu 和 Hwee Tou Ng 于 2002 年首次在事件抽取中引入最大熵分类器[9],用于事件元素的识别;David Ahn 2006 年结合 MegaM 和 Timbl 两种机器学习方法分别实现了事件抽取中事件类别识别和事件元素识别这两个主要步骤,在 ACE 英文语料上均取得了不错的效果[4]。但 Ahn 的方法由于将每个词作为一个实例来训练机器学习模型,引入了大量的反例,导致正反例严重不平衡;此外,事件类别的多元分类以及为每类事件元素单独构造多元分类器在语料规模较小的时候存在着一定的数据稀疏问题。

鉴于上述方法的不足,本文提出一种基于触发词扩展和二元分类相结合的识别方法进行事件类别的识别,多元分类模型的方法进行事件元素的识别,较好的避免了正反例不平衡和数据稀疏问题。

图 2 给出了本文事件抽取系统的系统框架图。

3　事件类别识别

事件触发词直接引发事件的产生,是决定事件类别的重要特征。本文提出基于触发词扩展和二元分类相结合的方法解决事件类别识别问题,分为候选事件的抽取和候选事件的分类两个主要步骤。

3.1　候选事件的抽取

本文将含有触发词的句子称为候选事件。事件触发词直接决定候选事件及其候选类别的获取。由于训练语料中触发词(种子触发词)数量有限,容易造成新事件的丢失。如:“他偏瘫在床”。假设“偏瘫”不是种子触发词,该句就不易被识别成事件。但“偏瘫”和“瘫痪”词义相近,本文使用哈工大信息检索研究室的《同义词词林(扩展版)》自动扩充种子触发词,尽可能多的覆盖各种类型事件的触发词。扩展后的触发词及其所在事件的类别,组成二元组对(trigger,type),如:(瘫痪,Life/ Injure)等,并构成“触发词—事件类别”二元对照表。据此,给出候选事件的抽取算法,如下:

Step1: 预处理所要分析的文章,包括分句和分词;

Step2: 针对每一个句子,查看组成它的词语是否在“触发词—事件类别”对照表中;

Step3: 若存在这样的词 w,则认为这个句子是一个候选事件,且事件触发词为 w,候选事件类别为触发词 w 所对应的类型。若该句子含有多个这样的词 w,则认为该句子中存在多个事件,该句子是由不同触发词 w 触发的不同类型的候选事件。

经过这样的抽取过程,不但可以获得大量的候选事件,而且还为每个候选事件规定了一个可能的候选类别,为后续的候选事件二元分类奠定了基础。

3.2　候选事件的分类

候选事件中存在大量不符合对应候选类别的事件。本文采用分类的方法挑选出真正的事件。由于每个候选事件仅拥有一个可能的候选类别,因此可将候选事件类别识别看作一个二元分类问题,即判断候选事件是否是满足候选类别的事件。

本文选取了词法、上下文、词典信息等三类语言学特征对候选事件进行描述,如表 3 所示。

表 3　事件类别识别的特征描述

Feature	Description
F_L: 词法特征	
Trigger	触发词本身
Trigger POS	触发词词性
F_C: 上下文特征	
Context Word POS	事件触发词左/右侧 p 个词语的词性信息
Context Head Word Type	事件触发词左/右侧 q 个实体的核心词的 type 信息
Context Head Word Subtype	事件触发词左/右侧 q 个实体的核心词的 subtype 信息
F_T: 词典信息特征	
Thesaurus Code	事件触发词对应的《同义词词林(扩展版)》第 m 层词义编码

4　事件元素识别

通过事件类别的确定,相应的就获得了该类事件的模板,即获得了要抽取的元素标签。由于事件元素是由触发词所在事件的 Entity、Time Expression、Value① 表示的,我们称其为候选事件元素。基于此,可将事件元素识别任务看成分类问

① 来自 ACE 标准标注结果,分别对应着 ACE 的三项标注任务: 实体识别、时间表达式识别和属性词识别。

题,转换为对文本中每个候选元素进行类别标签识别(包含“None”标签,表示不是事件元素),在后续工作中从候选事件元素中挑选出真正的元素。

4.1 多元分类策略

根据分类对象的不同,本文采用了以下三种多元分类策略:

1. M_{single}:为所有类别的事件构造一个候选元素多元分类器;

2. $M_{multi\text{-}type}$:为每类事件(Type)分别构造一个候选元素多元分类器;

3. $M_{multi\text{-}subtype}$:为每类子事件(Subtype)分别构造一个候选元素多元分类器。

其中,M_{single}策略训练实例最为充裕,训练最为充分;$M_{multi\text{-}type}$和 $M_{multi\text{-}subtype}$分别对该类/子类事件的实例进行训练,训练实例噪音较小。其中,Ahn 采用了 $M_{multi\text{-}type}$策略解决事件元素识别[4]。

4.2 特征选取

由于将事件元素识别看作分类任务,特征的选取和发现尤为关键。综合分析,本文选取词法、类别、上下文、句法结构等四类特征多角度的描述候选元素,进行元素标签的识别,如表 4 所示。

表 4 事件元素识别的特征描述①

Feature	Description
F_L:词法特征	
Trigger/ POS	触发词本身/触发词词性
F_T:类别特征	
Event Type/ Subtype	候选元素所在事件的类别/子类别
ETV Type/ Subtype	候选元素的 type/ Subtype
ETV Class	候选元素的 class
ETV Mention Type/ Subtype	候选元素的 mention Type/ Subtype
Head Word/ POS	候选元素的核心词及其词性
F_C:上下文特征	
Context Word	事件触发词左/右侧 p 个词语
Context Word POS	事件触发词左/右侧 p 个词语的 POS 信息
F_S:句法结构特征	
BA	ETV 在事件触发词前面还是后面,这是一个二值特征,标记为“B”代表前面,“A”代表后面
Trigger Parse	事件触发词与其父节点的句法关系
Head Word Parse	ETV 的 Head Word 与其父节点的句法关系
Path	从当前 ETV 到触发词的句法关系路径

其中,由于触发词间接决定了事件模板,而事件类别/子类别直接决定了事件模板,因此,触发词、事件类别和子类别对元素类别识别举足轻重;其次,候选元素的相关特征及其核心词特征体现了候选元素的核心语义,也很有意义;除此之外,是否是满足事件模板的元素和上下文信息有很大的关系,因此上下文的词语及其词性信息、句法结构信息是很重要的特征。

5 评价与性能分析

5.1 语料来源及评价方法

本文将 ACE 2005 中文语料作为实验数据,共 633 篇。随机抽取 473 篇为训练集,80 篇为开发集,80 篇为测试集。其中 ACE 评测提供的训练语料不但标注了 Entity、Time Expression、Value 及其核心词的各种属性,而且还详细标注了事件的各种组成要素,如:触发词,类别、元素等信息。本文采用了传统的 F 值的评价方法,对事件抽取的两个关键步骤——事件的类别识别和事件的元素识别,以及事件抽取系统进行了全面系统的评价。

5.2 性能分析

5.2.1 事件类别识别的实验结果与分析

本文提出的基于触发词扩展和二元分类相结合的方法与 Ahn 方法[4]的实验结果对比如表 5 所示。

在选择相同的特征进行模型训练和测试时,实验结果显示本文的方法明显好于 Ahn 的方法:

1. Ahn 的方法召回率很低。在二元分类判断是否是触发词时,Ahn 引入了大量的反例,正例反例比例为 1∶70,数据严重不平衡;而本文使用触发词构造候选事件,仅将候选事件中的触发词作为训练实例,有效缩减了训练实例中反例的个数,正例反例比例为 4∶7,数据较为平衡。

2. Ahn 的方法准确率不高。由于语料规模较小,分类类别较多(33 个类别),造成数据较为稀疏;

① 说明:ETV 代表 Entity、Value、Time Expression;ETV Type、EVT Subtype、ETV Class、ETV Mention type、ETV Mention subtype 信息来自于 ACE 的标准标注。其中句法结构特征使用了 LTP 提供的依存句法分析器。

表 5　对比实验结果

	Ahn's Method			Our Method		
	R	*P*	*F*	*R*	*P*	*F*
Development	43.06 %	58.29 %	**49.53 %**	57.14 %	64.22 %	**60.48 %**
Test	38.91 %	52.36 %	44.64 %	54.86 %	69.29 %	61.24 %

本文为每一个候选事件限定一个候选类别,进行二元分类,即判别这个事件的类别是否是候选类别,这种方法有效避免了语料规模小而带来的数据稀疏问题,大大提高了事件类别识别的准确率。

5.2.2　事件元素识别的实验结果与分析

针对提及的三种不同的事件元素分类策略在开发集上做了以下三组实验,结果如表 6 所示:

表 6　三种多元分类结果对比

M_{ethod}	*R*(%)	*P*(%)	*F*(%)
M_{single}	63.49	65.82	**64.64**
$M_{multi\text{-}type}$	63.83	64.81	64.32
$M_{multi\text{-}subtype}$	63.76	63.98	63.87

由实验结果可以看出,在选择相同的特征进行模型训练和测试时,M_{single}方法的实验结果最优:

1. M_{single}方法训练实例充裕,模型训练较为充分,分类效果较好,在开发集上最终的 *F* 值达到了 64.64 %。把实例按照事件类别和子类别分配之后,每一类(或子类)事件的训练实例大大减少,$M_{multi\text{-}type}$每类训练实例多则 1 000 个,少则十几个,训练非常不充分;$M_{multi\text{-}subtype}$的多元分类,由于事件子类别较多,每类的训练实例更少,数据更加稀疏,从而导致实验效果不好。

2. 通过观察,事件类别所对应的模板中的某些元素标签并不是完全独立于事件类别的。比如: Life/ Be-Born 类别事件中含有 Time, Place 元素标签,而在 Business/ Start-Org 事件中也含有 Time, Place 元素标签,且二者所表述的意义相同,上下文环境也很类似。因此按照事件类别/ 子类别训练多个多元分类器的方法反而减少了这些元素的训练实例,造成了数据稀疏。

5.2.3　事件抽取系统的性能分析

采用 *F* 值的评测方法,在 ACE2005 80 篇开发集和 80 篇测试集上的实验结果如表 7 所示:

表 7　事件抽取系统在开发集和测试集上的结果

Data	*R*(%)	*P*(%)	*F*(%)
Development	47.25	33.25	39.04
Test	46.38	37.05	41.20

分析实验结果,由于错误级联,虽然独立的事件类别识别和元素识别模块的 *F* 值都在 60 %以上,事件抽取系统的最终 *F* 值仅有 40 %左右。导致最终系统 *F* 值不高的原因有很多,比如特征提取不够全面,触发词扩展不够充分,预处理模块带来的一些噪音,ACE 语料本身存在的一些错误标注等等,但这也说明事件抽取工作还有很大的研究空间和研究价值。

6　结论与未来工作

本文实现了一个事件抽取系统,集事件类别识别、事件元素识别功能于一体。针对事件类别识别任务,文本通过采用《同义词词林(扩展版)》自动扩展事件触发词,生成候选事件及其候选类别;继而对候选事件结合词法特征、上下文特征、词典特征从不同的角度描述候选事件,进行二元分类,在 ACE2005 语料上进行实验并取得了 61.24 %的 *F* 值。实验表明: 由于扩展触发词的引入和候选事件的生成,有效解决了训练数据正反例不平衡问题以及数据稀疏问题。针对事件元素识别任务,本文将其看作分类问题,引入丰富有效的特征,如词法特征、类别特征、上下文特征、句法特征等,本文对比分析了基于最大熵的三种多元分类方法,在 ACE2005 语料上进行实验并取得了 66.90 %的 *F* 值。实验表明: 为所有事件类别的候选元素构造一个多元分类器的方法由于其训练数据较为充足,避免了其他两种分类方法带来的数据稀疏问题。

中文事件抽取技术还处于初级阶段,因此还有很广阔的研究空间。如: 如何进行领域的移植,如何发现新类型的事件等等,都将成为我们下一步的工作。

参考文献:

[1] Naomi Daniel, Dragomir Radev and Timothy Allison. Sub-event based Multi-document Summarization [A]. In: Proceedings of the HLT-NAACL Workshop on Text Summarization [C]. 2003. 9-16.

[2] Elena Filatova and Vasileios Hatzivassiloglou. Event-based Extractive summarization [A]. In: Proceedings of ACL Workshop on Summarization [C]. 2004. 104-111.

[3] Wenjie Li, Mingli Wu and Qin Lu. Extractive Summarization using Inter- and Intra- Event Relevance [A]. In: Proceedings of the 44th Annual Meeting of the Association for Computational Liguistics [C]. 2006. 369-376.

[4] David Ahn. The stages of event extraction [A]. In: Proceedings of the Workshop on Annotations and Reasoning about Time and Events [C]. 2006. 1-8.

[5] ACE (Automatic Content Extraction) Chinese Annotation Guidelines for Events. National Institute of Standards and Technology [R]. 2005.

[6] 姜吉发. 自由文本的信息抽取模式获取的研究[D]. 中国科学院博士学位论文, 2004: 1-18.

[7] Mihai Surdeanu, Sanda Harabagiu, John Williams, et al. Using Predicate-Argument Structures for Information Extraction [A]. In: Proceedings of ACL [C]. 2003. 8-15.

[8] Mihai Surdeanu and Sanda Harabagiu. Infrastructure for Open-Domain Information Extraction [A]. In: Proceedings of the Human Language Technology Conference [C]. 2002. 325-330.

[9] Hai Leong Chieu, Hwee Tou Ng. A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text [A]. In: Proceedings of the 18th National Conference on Artificial Intelligence [C]. 2002. 786-791.