音乐领域典型事件抽取方法研究[1]

丁效[2]，宋凡，秦兵，刘挺

哈尔滨工业大学计算机学院，信息检索研究中心，哈尔滨 150001

E-mail: xding@ir.hit.edu.cn

摘 要：事件抽取是信息抽取领域一个重要的研究方向。本文从音乐领域的事件抽取出发，通过领域事件词聚类的方法自动发现音乐领域具有代表性的事件，然后采用基于关键词与触发词相结合的过滤方法简化了事件类型的识别过程。在事件元素识别中，本文采用了基于最大熵的事件元素识别方法。在本文构建的语料库下，最终事件类型识别的平均 F 值达到 82.82%，事件元素识别的平均 F 值达到 75.79%。

关键词：事件抽取，事件类型发现，事件类型识别，事件元素识别

Research on Typical Event Extraction Method in the Field of Music

Ding Xiao, Song Fan, Qin Bing, Liu Ting

Research Center for Information Retrieval of Computer Science & Technology School, Harbin Institute of Technology, Harbin 150001

E-mail: xding@ir.hit.edu.cn

Abstract: Event extraction is an important research point in information extraction. This paper focuses on the music domain, and describes a method based on triggers clustering for event type discovering. Then we propose a method based on the filtering of keywords and triggers for event type recognition. For the event argument recognition, the method which is based on maximum entropy model is proposed in this paper. Evaluations on our corpus give a final F-score of 82.82% and 75.79% for type recognition and argument recognition.

Keywords: event extraction, event type detection, event type recognition, event argument recognition

1 引言

事件抽取是信息抽取研究中最具挑战性的任务之一，旨在把人们感兴趣的，用自然语言描述的事件以结构化的形式呈现出来，如什么人，什么地方，什么时间，做了什么事。事件抽取在多文档文摘[1]，自动文摘[2,3]，自动问答[4]和信息检索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（1987~1998）[5]，作为 ACE（Automatic Content Extraction）[6,7] 会议的前身，在上个世纪八、九十年代对信息抽取领域起到了很大的促进作用，事件抽取(Scenario Template)始终是这一会议的评测项目之一。ACE 也于 2005 年引入了事件抽取(Event Detection and Recognition、Event Mention Detection）评测任务。

目前的事件抽取方法采用信息抽取技术抽取预先定义的一种或者几种事件，然而不同领域的事件类型互不相同，这样的方法依赖于人工定义的事件类型，需要耗费大量的人工劳动，导致作为信息抽取关键技术的事件抽取缺乏足够的自适应性。从技术路线角度看，解决事件抽取问题的方法主要有两种：基于模板的方法与基于机器学习的方法。基于模板的主要通过手工或自动生成事件模板，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。例如 Yankova 的足

[1] 基金资助：自然科学基金 60975055, 60803093; 国家 863 项目 2008AA01Z144

[2] 作者简介：丁效（1985-），男，硕士研究生，信息抽取；宋凡（1984-），男，硕士研究生，信息抽取；秦兵（1968-），女，教授，博士生导师，信息抽取和多文档文摘；刘挺（1972-），男，教授，博士生导师，信息检索和自然语言处理。

球事件抽取系统[8]以及 Lee 的基于限定域 Ontology 的气象事件抽取系统[9]等等。基于机器学习的方法把主要的精力放在分类器的构建和特征的发现、选择上，把事件抽取问题看成分类问题，选择合适的特征使用分类器来完成。Chieu 和 Ng 于 2002 年首次在事件抽取中引入最大熵分类器[10]，用于事件抽取中事件元素的识别。

本文借鉴 ACE 中事件抽取的相关概念，并结合实际的需求做了相应的调整，将其转移到音乐领域上来，从音乐新闻资讯中抽取出需要的结构化信息。通过领域事件词聚类的方法自动发现音乐领域典型事件，对典型事件分别从语料的获取、标注、事件的定义、算法的应用都做了尝试，相同的方法可以平行的应用到其它的事件类型上。

音乐领域的事件抽取任务与 ACE 事件抽取任务大致相同，主要包括以下三个步骤：

1. 事件类型发现：事件类型是事件抽取任务的一个基础，本文不同于传统事件抽取之处就在于不是预先定义好事件类型体系，而是通过基于领域事件词聚类的方法自动发现事件类型；
2. 事件触发词及事件类别的识别：事件触发词是指引起事件发生的词，是决定事件类别的重要特征；
3. 事件元素的识别：事件的元素是指事件的参与者，本文为音乐领域的两个典型事件制定了模板，模板的每个槽值对应着事件的元素。

图 1-1 详细的表述了一个音乐领域事件的构成。其中，“举办”是该事件的触发词，所述事件类别为演唱会。该事件由四个元素组成，“周杰伦”、“2010 年 6 月 11 日”、“台北小巨蛋”、“周杰伦超时代演唱会”分别对应着该类（演唱会）事件模板中的四个角色标签，即：歌手、时间、地点以及演唱会名。

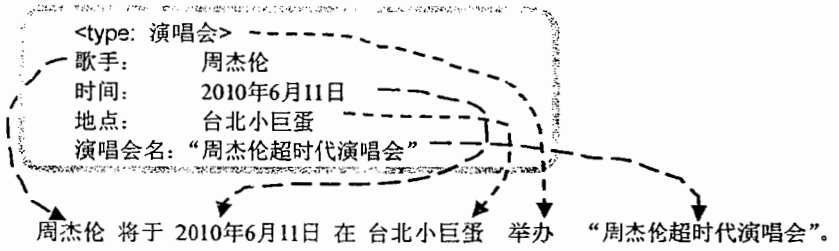

图 1-1 “演唱会”事件的基本组成要素

本文提出了基于聚类的自底向上的事件抽取方法。该方法依据谓语动词是对领域事件刻画的重要单元的特点，利用依存句法信息抽取领域事件词，利用《知网》(HowNet)对领域事件词进行聚类从而获取不同的事件类型，随后进行事件类型的识别及事件元素的抽取。和传统的事件抽取技术相比，该方法不需要预先定义事件类型，不需要先验的领域知识。因此，该方法是对领域移植的一个尝试，特别适用于知识和资源有限的领域。

2 基于领域事件词聚类的典型事件类型发现

目前学者们处理事件抽取的方法，一般是分为两个步骤：一是事件类型识别；二是事件元素识别。然而在实际应用时，会发现如果人工定义事件类型，则需要耗费大量的人力劳动，另外在某个特定领域，事件类型发现中需要先验知识，这种依赖使得不同领域特别是资源和知识有限领域的事件发现变得困难。因此，本文将自动发现事件类型作为音乐领域事件抽取的一个子任务。

事件类型发现分为两个步骤：第一领域事件词抽取，第二聚类领域事件词自动发现事件类型。

2.1 领域事件词抽取

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。而多数情况下事件触发词为动词，因此对领域事件词进行抽取是非常有必要的。

领域事件词抽取算法 DVE(Domain Verb Extraction)考虑领域语料中不同事件实例具有特定语义关系的动词，通过计算其在通用领域和目标领域中的分布情况抽取领域事件词，具体步骤如下所示：

Step 1:

对领域语料进行分句分词；使用依存句法分析器识别主谓关系 SBV 和动宾关系 VOB，主谓与动宾关系都关联的那个动词（V_i）即是作为本文抽取的候选领域事件词

Step 2:

根据动词 V_i 在领域语料和通用语料中的分布信息采用公式（2-1）计算其领域相关度 $DR(V_i)$，其中 $Freq_D(V_i)$ 和 $Freq_G(V_i)$ 分别表示 V_i 在领驭语料和通用语料中的出现概率

Step 3:

最后根据动词细分类（分为八大类：系动词、助动词、形式动词、趋向动词、补语动词、一般动词、名动词、副动词），将领域事件词再一次进行过滤，仅保留一般动词

$$DR(V_i) = \frac{Freq_D(V_i)}{Freq_G(V_i)} \tag{2-1}$$

其中依存句法分析器采用哈尔滨工业大学信息检索研究中心的依存句法分析模块 GParser。

2.2 基于领域事件词聚类的典型事件类型发现

动词是体现事件的最为重要的词汇单元，一系列具有相同含义和用法的动词体现同一类事件，因此采用基于领域事件词聚类的方法发现事件类型是可行的，本文提出的方法通过利用 HowNet 借助语义知识实现领域事件词聚类，从而发现事件类型。

本文提出的事件类型发现算法 ETDA(The Event Type Discovery Algorithm)不需要预先给定类型数量，而直接采用领域事件词聚类的方法获取事件类型。事件类型发现算法具体步骤如下所示：

Step 1:

构造候选事件实例<V_i, Obj>，其中 V_i 是抽取出来的领域事件词，Obj 是与领域事件词 V_i 构成动宾关系 VOB 的宾语

Step 2:

计算任意两个候选事件实例<V_i, Obj>相似度；若相似度大于阈值 0.6，则将其聚到一类

Step 3:

将那些没有聚到一起的候选实例归到未分类里面

在该算法中，候选事件实例相似度的计算由领域事件词间的相似度计算体现。本文中，利用语义相似度描述两个领域事件词 V_i 和 V_j，其相似度值 $Sim(V_i, V_j)$ 根据 HowNet 计算得到，$Sim(V_i, V_j)$ 由 V_i 和 V_j 相同“义原”的数量除以两者的“义原”总和实现归一化得到，如公式（2-2）所示。

$$Sim(V_i, V_j) = \frac{2N_s}{N_i + N_j} \tag{2-2}$$

其中 N_s 表示领域事件词 V_i 和 V_j 在 HowNet 概念意义 DEF(the concept definition in HowNet)中同样“义原”的数量，N_i 和 N_j 分别表示 V_i 和 V_j 概念定义中“义原”的数量。

通过该方法，在音乐领域语料上得到聚类后的两个典型音乐事件：演唱会事件和专辑事件。

2.3 典型事件类型识别

在自动发现了音乐领域两类典型事件后，下面要进行音乐领域典型事件类别识别工作。

由于处理的语料全部来源于音乐的新闻资讯，词的歧义性较少，而且现在所关注的两类事件都有明显的关键词标志，所以本文最终采用基于关键词与触发词过滤的方法，来形成候选事件。而对于候选事件的识别，首先进行事件元素的识别，然后看其组合是否符合本文定义的事件模板。

3 基于最大熵的事件元素识别

当完成候选事件识别之后，就要对候选事件中的众多实体中挑选正确的事件元素。例如句子："*周华健2008新年倒计时演唱会12月31日在上海举行，成龙、火炬手金晶等嘉宾捧场。*"通过关键词"演唱会"过滤为一个候选演唱会事件，接下来需要识别歌手周华健、成龙哪个才是真正的该事件的元素。该问题可以借助于机器学习分类器的方法来解决。

通过对大规模音乐领域语料的统计分析，发现一个句子是否能够成为本文所关注的演唱会及专辑事件，是要符合一定的事件模板，演唱会事件模板和专辑事件模板分别为：

(1) ***歌手[演唱会名字，时间，地点]***

其中，方括号内元素表示可选，该模板表示一个句子中首先必须包含歌手元素，然后还必须在演唱会名字、时间、地点三个要素中至少包含一个，才能算作是一个演唱会事件，否则应该放弃，不作为演唱会事件来看待。

(2) ***歌手[时间，专辑名]***

对于专辑事件，句子中必须包含歌手元素，然后还必须在时间、专辑名两个要素中至少包含一个才可视为专辑事件。

特征选择：

由于事件元素识别可以看作二元分类问题，本文为每类事件的每种元素训练一个二元分类器，这样一共有演唱会事件的歌手、时间、地点与专辑事件的歌手、时间、专辑6个二元分类器。

(1) 上下文特征（F_C:Context Features）

- 实体左侧 p 个词语
- 实体右侧 p 个词语
- 实体右侧 p 个词语的 POS 信息
- 实体右侧 p 个词语的 POS 信息

其中，p 为整数，且 $p \in [1,4]$

候选事件元素是否是真正的事件元素，是由它所在上下文中的语义所决定的，因此上下文信息对于事件元素的判定非常重要。本文选取两类上下文特征，上下文词语特征和上下文词语的词性特征。并把它们作为基本特征。

(2) 邻近实体特征（F_E : Neighbor Entites Features）

- 实体左侧 q 个实体的类型
- 实体右侧 q 个实体的类型

其中，q 为整数，且 $q \in [1,2]$

在识别候选事件元素的时候，邻近的实体信息对事件元素的识别也很有所帮助。例如：在演唱会事件中当歌手元素后面紧跟着时间或者地点元素的时候，它是一个真正的事件元素的概率就很大，而当它后面还是歌手元素的时候，往往就不是真正的事件元素。

(3) 规则特征（F_R : Rule Features）

- 所在子句中是否有触发词
- 该类型的元素在事件中是否唯一

如果与触发词在一个子句中的候选元素与触发词存在较强的语义关联性，则它是事件元素的可能性就相对较大。另外当某种类型的元素唯一时，它也很有可能就是事件元素。

(4) 句法结构特征（F_S :Syntax Features）

- 实体在句法树中父亲节点的词信息
- 实体在句法树中父亲节点的 POS 信息
- 实体结点与父亲结点的句法关系

大量的研究表明，句法结构特征能够很好的描述实例的特征及上下文的语境，所以根据依存句法分析的特点，本文选取候选实体如上的句法结构特征。

(5) 动词特征（F_V :Verb Features）

- 实体左侧最近的一个动词
- 实体右侧最近的一个动词
- 实体在句法树中最近的一个动词

触发词在判断事件元素的时候会起到决定性的作用。根据统计 ACE 中的触发词绝大多数都是动词，所以动词特征在判断事件元素的时候也有很大的影响力。

下面结合实例来详细描述各个特征，其中假设 p=2，q=1。例如事件：“*日前，F4 在日本横滨为 7 场巡回个唱揭开序幕。*”，考虑歌手元素“F4”，特征向量表示如图 3-2 所示。

F_C：前两个词和词性分别是“日前”“，”“nt”“wp”，后两个词和词性分别是“在”“日本横滨”“p”“Ns”。

F_E：前一个实体不存在，则标记为 null。后一个实体类型为地点 Ns。

F_R：所在子句有触发词“揭开”且该歌手类型元素在事件中唯一。

F_S：候选实体父亲结点的词是“揭开”，词性是“V”，与父亲结点的关系是 SBV。

F_V：前一个动词不存在，标记为 null，后一个动词为“揭开”，句法树中最近的动词也为“揭开”。

B1W=， B1T=wp B2W=日前 B2T=nt A1W=在 A1T=p A2W=日本横滨 A2T=Ns BE=null AE=Ns
tri=true uniq=true fatherT=v fatherW=揭开 fatherRe=SBV preVerb=null afterVerb=揭开 fatherVerb=揭开

图 3-2 候选事件元素特征集表示

4 实验结果与分析

4.1 语料来源与评价

本文使用 2008 年 08 月、2008 年 09 月、2008 年 10 月、2008 年 11 月、2009 年 03 月和 2009 年 04 月 6 个月的新浪音乐新闻资讯。最终标注语料 6000 句，拿出其中的 4000 句作为训练最大熵模型的训练集，1000 句作为最大熵模型的开发集，剩下 1000 句作为各种方法公共的测试集。训练语料中包含演唱会事件 1560 个，专辑事件 555 个；开发集中包含演唱会事件 335 个，专辑事件 155 个；而测试集中包含演唱会事件 422 个，专辑事件 160 个。

对于事件类型识别和事件元素识别的性能评价，本文采用了传统的 F 值的评价方法，定义如下：

(1) 事件类型的识别，定义如下：

$$F - socre = \frac{2PR}{P + R} \tag{4-1}$$

其中 P 为准确率，R 为召回率，分别定义为：

$$P = \frac{\text{识别正确的实例类型总数}}{\text{判别为有效实例总数}} \tag{4-2}$$

$$R = \frac{\text{识别正确的实例类型总数}}{\text{标准有效实例总数}} \tag{4-3}$$

(2) 事件元素的识别，定义如下：

$$F - socre = \frac{2PR}{P + R} \tag{4-4}$$

其中 P 为准确率，R 为召回率，分别定义为：

$$P = \frac{\text{识别正确的事件元素总数}}{\text{判别为事件元素总数}} \tag{4-5}$$

$$R = \frac{\text{识别正确的事件元素总数}}{\text{标准事件元素总数}} \tag{4-6}$$

4.2 实验结果与分析

经过实验验证，最终确定 p=2,q=1 并在最大熵训练的迭代次数为 100 时在开发集上达到最优。在分类的过程中，本文以 p=2 时的上下文特征为基本特征，系统为 baseline 系统。在这个系统中不断的加入新的特征，表 4-1 和表 4-2 列出了加入各种类型特征后的变化，这样可以清楚的观察到各种类型的特征在开发集上所起的作用。

表 4-1 演唱会事件各元素的二元分类结果

	演唱会歌手			演唱会时间			演唱会地点		
Feature	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
F_C	46.67	31.82	37.84	53.61	47.27	50.24	67.5	69.23	68.35
F_C+F_E	48.00	36.36	41.38	53.26	44.55	48.51	66.67	66.67	66.67
F_C+F_R	82.14	69.70	75.41	80.21	70.00	74.76	90.09	85.47	87.72
F_C+F_S	51.92	40.91	45.76	66.37	68.18	67.26	68.60	70.94	69.75
F_C+F_V	54.05	45.45	49.38	61.40	63.64	62.5	66.93	72.65	69.67
ALL	82.14	69.70	75.41	82.35	76.36	79.25	88.29	83.76	85.96

表 4-2 专辑事件各元素的二元分类结果

	专辑歌手			专辑时间			专辑		
Feature	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)	P(%)	R(%)	F(%)
F_C	60.23	43.80	50.72	67.42	57.69	62.18	72.65	85.86	78.70
F_C+F_E	71.11	52.89	60.66	75.58	62.50	68.42	73.55	89.90	80.91
F_C+F_R	89.80	72.73	80.37	90.00	77.88	83.51	93.14	95.96	94.53
F_C+F_S	66.29	48.76	56.19	76.19	61.54	68.08	74.58	88.89	81.11
F_C+F_V	66.02	56.20	60.71	79.00	75.96	77.45	73.60	92.93	82.14
ALL	91.84	74.38	82.19	90.72	84.62	87.56	93.14	95.96	94.53

从各种特征的对分类结果的贡献来看：

（1）两类事件中的规则特征都起了最重要的作用；

（2）其次是动词特征，因为事件元素左右动词可以为该事件元素提供很强的语义信息；

（3）句法特征，分析原因是由于现阶段本文仅仅选择了候选元素的父亲结点的相关信息作

为句法特征，相对来说还比较简单，而且与其他特征存在一定程度的重复，所以效果不是很明显；

（4）实体特征，对分类的结果贡献最小。

采用最大熵分类的方法在测试集上得到的结果如表 4-3 所示：

表 4-3 最大熵分类在测试集上的结果

评测对象		P(%)	R(%)	F(%)
事件类别识别	演唱会	87.43	66.51	75.55
	专辑	87.21	93.17	90.09
	平均	87.32	79.84	82.82
事件元素识别	演唱会	74.85	60.56	66.95
	专辑	82.43	86.95	84.63
	平均	78.64	73.76	75.79

5 结论与未来工作

本文针对音乐领域事件抽取的相关工作展开研究。对事件抽取的三个任务进行了探索性的尝试：（1）本文提出了基于领域事件词聚类的事件类型发现，解决了以往很少被人关注的一个事件抽取先行问题，同时，也是对领域迁移的一个尝试；（2）针对特定领域语料的基于触发词与关键词相结合的事件类型识别，也简化了以往事件类型识别的过程；（3）将事件元素识别看成分类问题，用最大熵模型将其解决。

在未来的工作中，将尝试选择更有效的句法特征以及其他特征；由于分类器的方法准确率要比基于模板的准确率低，因此将尝试基于模式匹配与最大熵相结合的方法抽取事件元素。

参 考 文 献

[1] Daniel N, Radev D, Allison T. Sub-event based multi-document summarization. In: Proceedings of the HLT-NAACL Workshop on Text Summarization. Edmonton, Canada, 2003: 9-16.

[2] Filatova E, Hatzivassiloglou V. Event-based extractive summarization. In: Proceedings of ACL Workshop on Summarization. Barcelona, Spain, 2004: 104-111.

[3] Li W J, Wu M L, Lu Q. Extractive summarization using inter- and intra- event relevance. In: Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics. Sydney, NSW, Australia, 2006: 369-376.

[4] Ahn D. The stages of event extraction. In: Proceedings of the workshop on annotations and reasoning about time and events. Sydney, NSW, Australia, 2006: 1-8.

[5] Chen Ji H. Language specific issue and feature exploration in Chinese event extraction. In: Proceedings of NAACL HLT 2009. Boulder, Colorado, 2009: 209-212.

[6] http://www.nist.gov/speech/tests/ace/

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

[8] Yankova M. Focusing on scenario recognition in information extraction. In: Proc. EACL. Budapest, Hungary, 2003: 41-48.

[9] Lee C S, Chen Y J, Jian Z W. Ontology-based fuzzy event extraction agent for Chinese e-news summarization. *Expert Systems with Applications*, 2003, 25(3): 431-447.

[10] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text. In: Proceedings of the 18th National Conference on Artificial Intelligence. Edmonton, Alberta, Canada, 2002: 786-791.