基于跨实体推理的事件抽取方法研究*

马彬，洪宇，杨雪蓉，姚建民，朱巧明
（苏州大学江苏省计算机信息处理重点实验室，苏州，215006）

摘　要：事件抽取是信息抽取领域一个重要的研究方向，旨在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。当前的许多研究工作都采用了基于直推式学习的方法，如跨事件推理的事件抽取研究。本文提出了一种新的基于直推式学习算法的事件抽取方法，跨实体推理(cross-entity inference)事件抽取。与以往的直推式事件抽取方法不同，跨实体推理事件抽取的核心是充分利用实体类型的一致性特征。将本文的方法应用于传统句子级别的 ACE 事件抽取任务中，实验结果显示，触发词的识别和事件元素/角色的识别性能分别获得了 8.6%和 11.8%的提高。

关键词：直推式学习；跨实体；事件抽取；信息抽取；自然语言处理

Using Cross-Entity Inference to Improve Event Extraction

Bin Ma, Yu Hong, Xuerong Yang, Jianmin Yao, Qiaoming Zhu
(Provincial Key Laboratory of Computer Information Processing Technology
Soochow University, Suzhou, China, 215006)

Abstract：Event extraction is the task of detecting certain specified types of events that are mentioned in the source language data. The state-of-the-art research on the task is transductive inference (e.g. cross-event inference). In this paper, we propose a new method of event extraction by well using cross-entity inference. In contrast to previous inference methods, we regard entity type consistency as key feature to predict event mentions. We adopt this inference method to improve the traditional sentence-level event extraction system. Experiments show that we can get 8.6% gain in trigger (event) identification, and more than 11.8% gain for argument (role) classification in ACE event extraction.

Key words: Transductive Inference; Cross-Entity; Event Extraction; Information Extraction; Natural Language Processing

1 引言

根据 ACE(Automatic Content Extraction)事件抽取的任务定义，事件抽取包含三个主要步骤：事件类型识别、事件元素识别和事件角色识别[1]。近期的许多事件抽取工作都采用了基于直推式算法的思想，如跨文档(cross-document)事件抽取、跨句子(cross-sentence)事件抽取和跨事件(cross-event)事件抽取。直推式算法的核心思想是根据已知实例的信息，推断同类型实例的未知属性信息。比如，对于某一候选事件，在跨事件的事件抽取方法中，可以根据同一篇文章中共现的相关事件信息判断该候选事件的类型和元素/角色信息。例如，给定句子(1)：

(1) *He **left** the company.*（译文：“*他离开了公司*。”）

从句子的局部信息很难判定该事件为 Transport 事件（即“地点转移”事件，意为某人离开某地）还是 End-Position 事件（即“职位终止”事件，意为某人从特定岗位辞职）。跨事件事件抽取根据同一篇文章中的“*Then he went shopping*”（译文：“然后他去购物”）事件信息将(1)正确的判定为 Transport 事件。

*基金项目：本课题得到国家自然科学基金（No. 61003152, 61272259, 61272260, 90920004, 61373097），教育部博士学科点专项基金（No. 20103201110021），江苏省自然科学基金（No. BK2011282），江苏省高校自然科学基金重大项目（No. 11KJA520003）以及苏州市自然科学基金（No. SYG201030, SH201212）资助。

作者简介：马彬（1989-），男，硕士，研究方向为事件抽取、话题检测和信息抽取；洪宇（1978-），男，副教授，通信作者（tianxianer@gmail.com），主要研究方向为话题检测、信息检索和信息抽取；杨雪蓉（1990-），女，硕士，主要研究方向为事件关系检测和信息抽取；姚建民（1971-），男，教授，研究方向为数据挖掘和机器翻译；朱巧明（1963-），男，教授，主要研究方向为嵌入式和自然语言处理。

从句子(1)的判断过程中可以看出，基于直推式学习算法的事件抽取很大程度上依赖于同类型实例的先验知识。然而，通常情况下，这种先验知识是缺省的，并且很难获得。在利用跨事件的方法将句子(1)判定为 Transport 事件类型的例子中，事件之间的关系是非常重要的先验知识，而事件关系抽取本身就是信息抽取领域中很难的一项任务。所以，跨事件事件抽取很大程度上会受到不相关的事件信息干扰。

本文提出一种新型直推式学习算法，即跨实体推理，解决事件抽取问题。跨实体事件抽取旨在利用实体与实体的一致性，在同类实体中实现直推式推理，即利用已知实体参与的事件属性，推理同类实体参与的事件属性。本文方法的提出源于以下文本现象：相同类型的实体会经常出现在类似的事件类型中。因此，可以利用实体类型的一致性，进行事件类型的推理。如，句子(2)：

(2) ***Obama*** *beats McCain.*（译文：“*奥巴马击败麦凯恩。*”）

其中，触发词“*beat*”（译文：“击败”）能够触发两种类型的事件：Elect（选举）和 Attack（袭击）。如果已经将句子“*Bush beats McCain*”（译文：“布什击败麦凯恩”）正确判定为“选举”事件，由于“奥巴马”和“布什”属于同种实体类型且具有相同背景知识（都是美国总统候选人），跨实体推理可借助“布什”参与“选举”事件，推理“奥巴马”参与的“击败”事件也为“选举”类型事件。

实体类型的一致性特征不仅可以用来预测事件类型，同样可以用在事件抽取的各项工作中：

- → 事件类型识别：相同类型的实体经常出现在相同或者相似类型的事件中，且使用相同或者类似的触发词。
- → 事件元素识别：与某一类型的实体共现的事件参与者，往往是同类型实体，甚至是相同的实体。
- → 事件角色识别：同类型的元素，在相同或者相似类型的事件中扮演相同的角色。

因此，可以利用实体在事件抽取中表现出的特点，通过以下步骤完成跨实体事件抽取的推理过程：

步骤一：将与句子中某实体类型相同的实体集合作为先验知识，预测事件类型并识别事件触发词。
步骤二：根据获得的实体类型、事件类型和触发词识别事件元素列表。
步骤三：通过实体类型、事件类型、触发词和元素列表，决定事件元素的角色信息。

基于以上步骤，本文提出一种隐式的跨实体事件抽取方法。首先，利用句子中的实体作为查询词，从大规模的语言资源中检索相关文档，进而通过相关文档构建实体类型描述。然后，通过计算该实体与实体类型描述的相似度判断实体的类型。最后，利用训练语料中的先验事件属性知识，并结合实体类型的一致性特征，在测试集（候选事件）中按照以上步骤，进行跨实体事件抽取的推理过程。

与其它基于直推式学习的事件抽取方法不同的是，跨实体事件抽取的核心是有效利用实体关系信息。因此，推理过程具有以下优点：1）避免了大量错误的外部信息或先验知识对系统的影响（因为实体类型的一致性较为容易判断），因为可以利用从互联网上获得的相关文档信息判断实体类型的一致性。2）丰富的外部知识（同种类型的实体信息较为丰富），因为任何实体都存在大量的近义词和同义词。

2 任务描述

事件抽取是自动内容抽取（Automatic Content Extraction，简称 ACE）的子任务。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。事件抽取要求从含有事件信息的非结构化源文本中，自动识别和抽出含有事件类型、事件元素和事件角色信息的结构化信息。为了便于内容理解，此处给出 ACE 事件抽取中相关术语的定义：

- **实体(Entity)**：属于某个语义类别的对象或对象集合。
- **实体描述(Entity mention)**：包含实体的短语（通常情况下是名词短语）。
- **事件触发词(Event trigger)**：引发事件发生的核心词（ACE 中触发词主要为动词或者名词）。

- **事件元素(Event arguments)：**事件的参与者，是组成事件的核心部分。
- **事件角色/元素角色(Argument roles)：**事件参与者与事件的关系。
- **事件描述(Event mention)：**包含事件触发词和事件参与者的短语或者句子。

ACE2005 定义了 8 种事件类别以及 33 种子类别。本文简单认定为 33 种事件类型，不考虑类别的层次关系。事件抽取正确识别一个事件当且仅当识别出所有的事件信息：事件类型、触发词、事件参与者和事件角色。以句子(3)为例：

(3) *60-year-old Mohammed al-Biyari was killed in his home near Jabaliya refugee camp by the rocket.*[1]
（译文：“*60 岁的默罕默德在贾巴利亚难民营附近的家中被火箭弹射杀*。”）

事件抽取应该正确的识别该事件为 Die 类型的事件，并且识别出相应的事件元素和角色[2]。见表 1。

表 1 事件抽取示例

事件类型	***Die***	
触发词	***killed***	
事件元素	***rocket***	角色=***Instrument***
	60-year-old Mohammed al-Biyari	角色=***Victim***
	his home near Jabaliya refugee camp	角色=***Place***

事件抽取很大程度上依赖于信息抽取领域其它方面的工作，如实体识别，实体共指研究和分类等。然而，实体识别本身就是 ACE 的一项子任务，本文不作为重点，在事件抽取过程中直接使用该部分的 ACE 标注结果。

3 相关工作

目前，针对事件抽取的研究主要集中于抽取模型的构建和直推式推理机制的设计，下面分别予以介绍。

Finkel 等（2005）[3]通过吉布斯采样，在因素概率模型中将蒙特卡洛（Monte Carlo）方法用于近似推理，并且通过模拟退火算法，取代了序列模型（HMM，CMM 和 CRF）中的维特比编码算法。在保留简单推理的情况下，融入了全局信息进行事件属性的独立抽取。Maslennikov 等（2007）[4]采用模式匹配的推理模型，其中使用句法树和局部句法依存关系知识，并结合上下文信息，对收敛于特定句子中的事件进行分类抽取。Patwardhan 等（2007，2009）[5][6]提出了双模型事件抽取系统：一个模型用于句子级别的事件识别，其通过提供概率评估，判断一个句子是否是领域相关的事件；另外一个模型用于识别候选的角色信息。系统允许两个模型在参考了每个短语的局部上下文信息和间接的全局信息后，共同对事件触发词和事件元素的识别给予判定。

近期的许多事件抽取工作都采用了基于直推式算法的思想，如跨句子(cross-sentence)、跨事件(cross-event)和跨篇章(cross-document)的事件抽取。直推式算法的核心思想是：根据已知实例的信息推断同类型实例中的未知属性信息。Ji 等（2008）[7]沿用 Yarowsky（1995）[8]的“单片断单语义”假设，将事件抽取的范围从单文档引申到话题相关的文档集合中，并且使用基于规则的方法，解决了跨句子和跨文档的触发词分类问题。将全局信息与局部信息相结合，在事件类型识别和事件元素识别中都获得了较大的性能提升。Gupta 等（2009）[9]通过跨事件的推理方法进行事件抽取，其基本思想是同类事件具有相同的属性（类型、元素和角色）。Gupta 借助语义特征对事件进行分类，在同类事件之间实现事件属性的推理。Liao 等（2010）[10]提出了文档级别的跨事件推理方法。与 Gupta 的工作相比，Liao 充分利用了相关事件的内容信息和事件类型一致性等特征，在事件预测和解决事件歧义性方面起到了很好的效果。

[1] 来源于 ACE2005 语料库中的文件“AFP_ENG_20030305.0918”

表 2 跨实体推理实例

(5)	事件类型	***Attack***	
	触发词	***war***	
	事件元素	***American***	角色=***Attacker***
		Saddam Hussein	角色=***Target***
(6)	事件类型	***Attack***	
	触发词	***Torture***	
	事件元素	***Bush***	角色=***Attacker***
		...Qaeda chief ...	角色=***Target***

4 动机

目前，基于直推式学习的事件抽取方法，都旨在解决因局部信息稀少造成的事件错标和漏标现象，核心思想是通过挖掘与事件相关的全局信息，将全局信息作为预测未知事件及其属性的先验知识，并结合句子的局部信息进行触发词抽取、事件元素识别以及角色标注，主要方法有跨文档事件抽取、跨事件事件抽取等。

通过分析句子级别的事件抽取发现，实体信息是事件抽取任务中主要信息来源，并且实体中蕴含的背景知识将进一步推动事件抽取的进行。如，利用“*Vesuvius*”是个活火山的背景知识，将和“*Vesuvius*”在一个句子中共现的词语“*erupt*”识别为事件“*volcanic eruption*”（火山爆发）的触发词而不是事件“*spotty rash*”（斑点皮疹）的触发词。

通过以上分析发现，实体一致性特征会对事件抽取研究带来帮助，但是如何将实体信息应用于事件抽取中是本文将要解决的难点。本文的主要目的是利用已知实体对应的事件信息，推理具有相同背景知识的实体所对应的未知事件信息。如，具有相同的背景知识的实体 a 和 b，若已知实体 a 事件中充当某个角色，可以根据实体的一致性特征，推断实体 b 在很大概率上也会在相同或类似的事件中充当相同角色。考虑下面两个事件[2]：

(4) ***American*** *case for* ***war*** *against* ***Saddam***.（译文：“*美国主张发动针对萨达姆的战争。*”）

(5) ***Bush*** *should* ***torture*** *the* ***al Qaeda chief operations officer***.（译文：“*布什应该惩治基地首领。*”）

实体“*Saddam*”和“*Qaeda chief*”具有相同的背景知识（恐怖分子首领），并且都充当 Attack 事件的 Target 角色。所以，在事件抽取中，可以利用已知“*Saddam*”在（4）中的角色信息推测“*Qaeda chief*”在（5）中的角色情况（见表 2）。

综上所述，跨实体事件抽取的思想可归纳为基于以下假设：具有相同类型的实体会出现在类似的事件中，并且在事件中充当类似的角色。下一节，通过对 ACE 语料中相关数据的统计，验证以上假设。

表 3 与实体类型 Population-Center 共现概率大于 0.05 的事件类型

事件类型	条件概率	频度
Transport	**0.368**	**197**
Attack	**0.295**	**158**
Meet	**0.073**	**39**
Die	**0.069**	**37**

[2] 来源于ACE2005语料库中文件“CNN_CF_20030305.1900.00-1”和文件“CNN_CF_20030303.1900.06-1”

4.1 实体一致性和分布情况

通过对 ACE 语料的统计发现，相同类型的实体具有强烈的一致性：某类型的实体在某事件中出现，那么该类型的其它实体会很大概率的也出现在该事件或者该类型的事件中，并且会使用相同的触发词表征事件的发生。为了验证以上假设，本文统计了部分实体类型在 33 种事件类型中的概率分布情况（见图 1）。

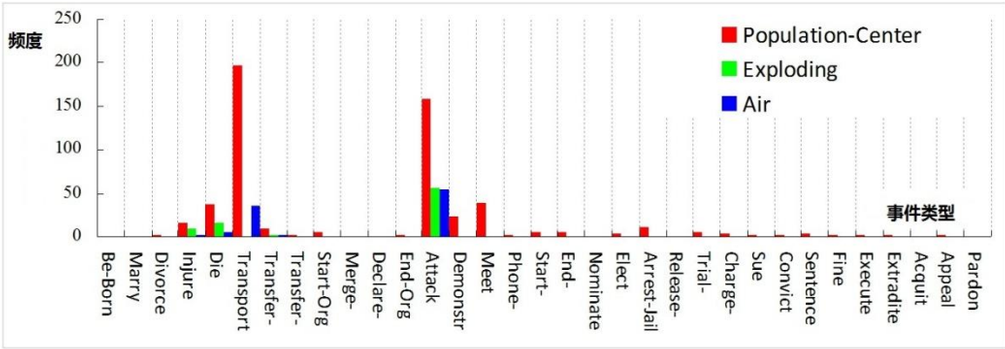

图 1 实体类型在 ACE 定义的 33 种事件类型中出现的条件概率

（以实体类型 Population-Center，Exploding 和 Air 为例）

ACE 事件抽取任务定义了 33 种事件类型，43 种实体类型，因此有 33*43 种实体-事件组合。由统计信息发现，只有小部分实体-事件组合出现的频率较高。如实体类型 Population-Center 只在 4 种事件类型中出现的概率大于 0.05。从表 3 中看出只有 Transport 事件和 Attack 事件与实体 Population-Center 会经常同时出现。

对于多数实体类型，与其高频共现的事件类型数目非常稀少。Air 类型的实体只与 5 种事件类型（Attack，Transport，Die，Transfer-Ownership 和 Injure）共现；Exploding 类型实体只和 4 种事件类型共同出现（见图 1），而且与其共现概率大于 0.05 的事件类型数目都只有一个或者两个。

表 4 ACE 语料中实体类型和事件类型共现分布:情况

	事件类型数<=5	5<事件类型数<=10	事件类型数>10
频度 ＞0	**24**	**7**	**12**
频度 >10	**37**	**4**	**2**
频度 >50	**41**	**1**	**1**

表 4 给出了 ACE 语料中实体类型和事件类型共现的概率分布情况。分析表可以发现：有 43 种实体类型在所有事件中出现的次数超过 10 次，其中的 37 种实体类型，与其共现的事件类型数都少于或者等于 5 次，只有 2 种实体类型会分布在超过 10 种事件类型中；当实体类型在 33 种事件类型中出现的频率超过 50 时，有 41 种实体类型（占实体类型总数的 95%）只分布在少于或等于 5 种事件类型中。以上分布现象表明：对于大多数类型的实体，只会出现在极少数的事件类型中。这种分布现象有助于事件抽取过程中，借助已知的实体信息促进事件类型的确定和触发词的识别。

另外，根据背景知识的相似度，将 ACE 中的实体类型更细粒度的划分为子实体类型。子实体类型的划分更能体现出实体类型与事件类型的对应性分布特点。比如，实体类型 Air 可以分为 Fighter plane，Spacecraft，Civil aviation 和 Private plane 四种子类型，其中 Fighter plane 类型的实体全部出现在 Attack 事件中，而其它三种子类型的实体只会在 Transport 事件中出现（见表 5）。实体子类型的一致性特征将有助于提高事件类型识别的准确率。

表 5 与 Air 实体类型共现的事件类型分布

Air (实体类型)	
Attack 事件	***Fighter plane*** (实体子类型 1): "*MiGs*" "*enemy planes*" "*warplanes*" "*allied aircraft*" "*U.S. jets*" "*a-10 tank killer*" "*b-1 bomber*" "*a-10 warthog*" "*f-14 aircraft*" "*apache helicopter*"
Transport 事件	***Spacecraft*** (实体子类型 2): "*russian soyuz capsule*" "*soyuz*"
	Civil aviation (实体子类型 3): "*airliners*" "*the airport*" "*Hooters Air executive*"
	Private plane (实体子类型 4): "*Marine One*" "*commercial flight*" "*private plane*"

4.2 角色一致性和分布情况

ACE 中事件和角色的共现情况呈现类似于 4.1 中描述的规律：同类型的实体在事件中充当相同的角色，尤其是在相同类型的事件中。如实体类型 Population-Center 的实体在 ACE 中只会表现出 Place，Destination，Origin 和 Entity4 种事件角色类型，对应出现的频率分别为 0.615，0.289，0.093 和 0.002（见图 2），因此，Population-Center 类型的实体主要表现出 Place 角色（在 Transport 事件中）和 Destination 角色（在 Attack 事件中）。同时，从图 2 中还可以看出，实体类型 Exploding 只表现出两种事件角色：Instrument（0.986）和 Artifact（0.014），并且主要表现出 Instrument 角色类型。

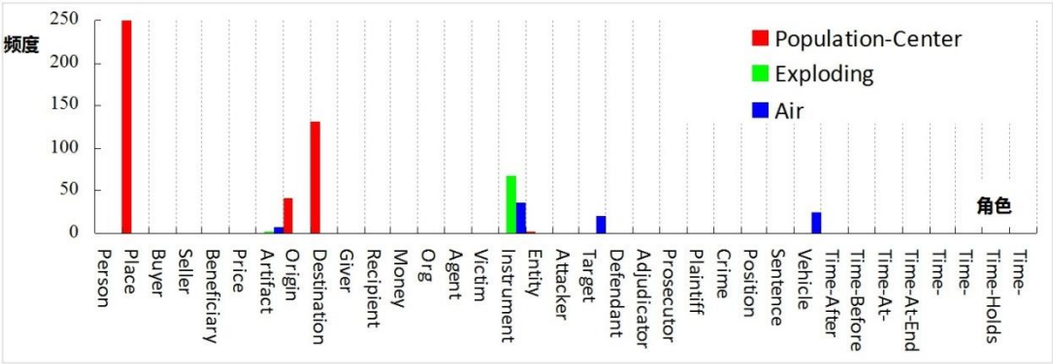

图 2 实体类型在 ACE 定义的 34 种角色类型中出现的条件概率
（以实体类型 Population-Center，Exploding 和 Air 为例）

表 6 给出了 ACE 语料中实体类型和角色类型的共现分布情况。分析表发现，在所有角色类型中出现频度大于 10 的实体类型一共有 43 种，其中的 38 种对应的角色类型数都小于或者等于 5，只有 2 种实体类型对应的角色类型数目大于 10。上述分布情况表明：同类型的实体，在事件中只会表现出特定数目的角色类型。这种分布现象可以通过跨实体的方法，促进事件抽取中角色的识别。

表 6 ACE 语料中角色类型与实体类型共现分布情况

	角色类型数<=5	5<角色类型数<=10	角色类型数>10
频度 > 0	**32**	**5**	**6**
频度 >10	**38**	**3**	**2**
频度 >50	**42**	**1**	**0**

5 跨实体事件抽取方法

本文针对句子一级的事件属性抽取进行推理，主要推理框架如图 3 所示。其中，实体类的划分侧重检测具有一致背景的实体，从而辅助基于实体类别一致性的跨实体推理。推理过程通过触发词识别、事件类型识别、事件元素识别、角色识别和可选事件识别，逐步完成特定事件的属性抽取。在此基础上，跨实体推理基于支持向量机（SVM）分类器实现事件属性的自动判定：

- → 事件元素分类器：划分事件描述（即候选句）中触发词所涉及的事件元素[3]。
- → 事件角色分类器：划分事件元素的角色信息。
- → 可选事件分类器：给定实体类型，触发词，事件类型和事件元素等特征，判断句子是否含有可选事件描述。

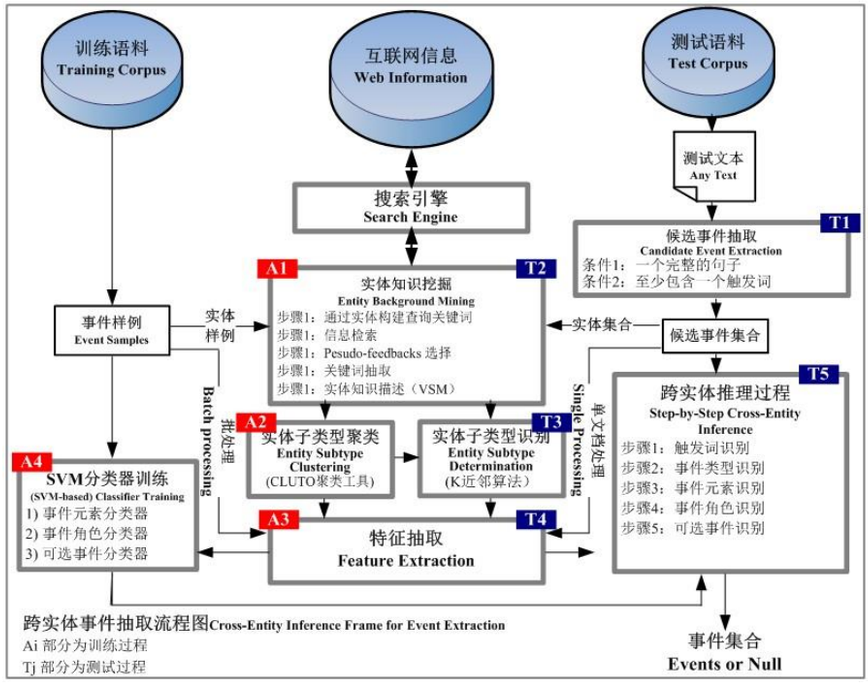

图 3 跨实体推理的事件抽取流程图（包括训练和测试）

下面分别对跨实体推理的各个组成部分予以介绍。

➢ 事件类型推理

给定一个候选事件描述（候选句），跨实体推理通过如下方式判定触发词和事件类型：首先，选择候选事件描述中的某个实体，判断其实体类型（假设类型为 i）；然后，将触发词列表与该事件描述中的非实体描述部分进行匹配，如果匹配出某个触发词（假设触发词为 j），则把与实体类型 i 共现频率最高，且触发词为 j 的事件类型，判定为候选事件的事件类型。

➢ 实体类划分与检测

实体类划分利用聚类技术（CLUTO 工具）[4]将 ACE 语料的实体类型按照背景知识的异同划分成不同的子类型。比如 Air 实体类型对应 Fighter plane，Spacecraft，Civil aviation，Private plane 四种实体子类型（见表 5）。将同一类型的实体集合按照背景的异同划分为实体子类型，相同子类型的实体具有更强的

[3] 一个句子中有可能含有不止一个事件，所以有必要先识别出触发词然后在识别对应的事件元素信息。

[4] http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA439508。

一致性，更有益于跨实体事件抽取的推理。

划分过程中，首先收集 ACE 语料中每种实体类型对应的实体集合，然后针对每一个实体，从互联网检索最相关的 50 篇文档，通过计算 50 篇文档中每个词语的 TFIDF 值选择权重最高的 50 个关键词语作为该实体的背景描述。将每个实体的背景特征表示为向量（向量空间模型，VSM），使用聚类工具（CLUTO）对实体集合进行聚类，获得的每一个聚类为一个实体类。

- **事件元素划分**

事件元素划分过程借助事件类型，限定候选事件元素出现的范围。通过 SVM 分类器识别候选事件描述中的事件元素信息，分类器的每一维特征由以下三项组成：

- → 事件类型；
- → 已知元素的实体类型；
- → 二元指示器，如果该事件类型中含有其它共现的实体类型，则为 1，否则为 0。

其它特征见表 7，如参数是否和触发词出现在同一个句子中等。

- **事件角色划分**

通过事件元素分类器获得的事件元素信息，为事件角色的识别提供了丰富的上下文信息。如当“市民”（事件元素一）和“恐怖分子”（事件元素二）出现在同一个事件中时，有很大概率将元素一标注为“受害者”角色。事件类型在事件角色标注过程中同样起到很重要的作用，使得事件角色的预测会更加精确。

此外，相同类型的实体在相同或者相似类型的事件中，往往会充当相同的角色，特别是当与其共现的事件元素类型相似的情况下。所以，实体子类型、事件类型和事件元素都能为事件角色的判定，提供丰富的上下文信息，可以作为角色分类的主要特征。基于 SVM 的事件角色分类器包含如下特征：

- → 特征 1 和特征 2（见表 4）；
- → 二元指示器，指示参与待测事件的元素的实体类型（因为有 266 种实体类型，所以此处对应 266 维二元特征）。

- **可选（Reportable）事件识别**

事件抽取任务中，存在两种影响事件抽取性能的问题：其一，触发词表中存在误导性极强的通用词语，如指代词“it”，“this”，“what”等。这些词语只在少数的事件中充当触发词（作为行为或状态的指代），然而一旦出现在触发词列表中，会引入大量的噪声，给后续事件类型的甄别产生误导；其二，有些事件元素会被标注成多个事件角色，根据 ACE 事件抽取纲要中的事件抽取任务定义，一个事件中的每个事件元素只能被标注一个特定的事件角色。为解决上述两个问题，需要对触发词和事件角色进行排序，移除置信度较低的触发词和角色信息。本文采用置信度系数将正确的触发词和角色与错误的区分开。置信度系数由两方面组成，一是在事件描述的特定范围内触发词（或者角色）的出现频率；二是在整个训练语料中触发词（或者角色）的出现频率。将两部分的得分进行加权，具体计算公式如下：

$$Conf(t) = \alpha \cdot norm.tf(t) + \beta \cdot \sigma(t) \qquad \text{公式(1)}$$

其中，*Conf(t)*表示词语 *t* 作为事件触发词（或者事件角色）的置信度值，α 和 β 为加权系数，*norm.tf(t)* 表示词语 *t* 在测试语料中的归一化词频（出现的频度除以文章中频度最大的词语对应的频度值），反应了词语在当前文档的重要程度；$\sigma(t)$ 表示词语 *t* 在训练文本中作为事件触发词（或者事件角色）的方差信息，表征词语在不同文档中作为触发词（或者事件角色）的稳定程度。$\sigma(t)$ 的计算公式如下：

$$\sigma(t)=\frac{1}{N}\sum_{i=1}^{N}(\frac{n_i}{N_i}-\overline{p})^2 \qquad \text{公式(2)}$$

其中，N 为训练语料中文档数目，N_i 为词语 t 在文本 i 中出现的频度，n_i 为词语在文本 i 中作为触发词（或者事件角色）的频度，$\overline{p}$ 为所有训练文本对应的的 n_i/N_i 平均值。

本课题将上述置信度作为评判可选事件的特征，借助 SVM 的分类器识别和屏蔽噪声事件，分类器采用如下三项特征：

- → 事件类型（确定事件描述的领域信息）；
- → 每种事件领域内特定触发词的置信度值；
- → 每种事件领域内特定角色的置信度值。

表 7 SVM 分类器的特征集合

事件元素分类器 **Argument Classifier**
特征 1: 事件类型 an event type
特征 2: 实体子类型 an entity subtype
特征 3: 共现的实体子类型 entity-subtype co-occurrence in domain
特征 4: 和触发词的距离 distance to trigger
特征 5: 和其它事件元素的距离 distances to other arguments
特征 6: 是否和触发词在句子中共现 co-occurrence with trigger in clause
事件角色分类器 **Role Classifier**
特征 1* 和 *特征 2
特征 7: 事件元素的实体子类型 entity-subtypes of arguments
可选事件分类器 **Reportable-Event Classifier**
特征 1
特征 8: 触发词置信度值 confidence coefficient of trigger in domain
特征 9: 事件角色置信度值 confidence coefficient of role in domain

6 实验

本文继承了 Liao（2010）的事件抽取评测方法。选取 ACE2005 英文语料中的 40 篇新闻文本作为测试集合，10 篇作为开发集合（用于分类器训练过程中的参数调整），剩余的 549 篇新闻文本作为训练集。

为了能与跨事件推理的事件抽取方法（Liao，2010）进行对比，本文同样对 10 组数据（每组 40 篇测试文本）进行测试，并记录最优值，最差值和平均值（见表 8）。系统采用准确率（P 值），召回率（R 值）和 F 值作为评价指标。同时，表 8 给出了两组（标注者 A 和标注者 B）人工标注的性能，其中标注者 A 明确和熟悉事件抽取任务，标注者 B 则不是。

6.1 实验结果

在开发集合上对公式(1)中加权系数进行调整，使其获得最佳结果，随后应用到测试集中，加权系数取值分别为：α= 0.05、β= 0.95。β 取值较大原因在于：由于 $\sigma(t)$ 的值较小，因此为了凸显方差信息的有效性，β 的取值较大。实验结果如表 8 所示。通过分析实验结果可以看出，跨实体事件抽取在触发词识别部分，F 值获得了 8.59%的提高，事件元素的识别提高了 11.86%，事件角色识别提升了 11.9 个百分点。

相比于跨事件的抽取方法，本文的方法在事件元素识别和角色标注部分分别提升了 2.87%和 3.81%，值得指出的是，本系统的最差性能仍然优于跨事件事件抽取的性能。

跨实体事件抽取对触发词识别的效果仍然不是很理想，较低的召回率严重影响了 F 值的提高。在实验中，系统选择至少包含一个实体描述的句子作为候选事件描述，然而 ACE 语料的许多事件是不含实体描述的，所以造成了部分候选事件的丢失。

另外，通过与两位标注者的结果对比可以发现：系统性能与了解事件抽取任务的标注者 A 的标注结果较为接近，且整体性能明显优于从未从事事件抽取工作的标注者 B 的实验结果。对比的结果有效说明了本文方法的正确性。比较标注者 A 和标注者 B 的结果发现，明确事件抽取任务能有效促进事件触发词的正确识别和事件角色的确定，但是对促进事件元素的识别作用不明显，即用户在不参考事件类型和实体类型的情况下仍然能正确识别出事件元素信息。因此，在以后工作中，在触发词识别和事件角色识别部分，可以尝试进一步挖掘事件抽取任务语料中数据的分布特点，以达到充分利用事件类型、实体类型和事件角色相互间的蕴含关系的目的。

表 8 实验结果

系统性能 /人工结果	**触发词识别 (%)**			**事件元素识别 (%)**			**事件角色识别(%)**		
	P	R	F	P	R	F	P	R	F
句子级别的 baseline	67.56	53.54	59.74	46.45	37.15	41.29	41.02	32.81	36.46
跨事件推理	**68.71**	**68.87**	**68.79**	**50.85**	**49.72**	**50.28**	**45.06**	**44.05**	**44.55**
跨实体推理（最优值）	73.4	66.2	69.61	56.96	55.1	56	49.3	46.59	47.9
跨实体推理（最差值）	71.3	64.17	66.1	51.28	50.3	50.78	46.3	44.3	45.28
跨实体推理（平均值）	**72.9**	**64.3**	**68.33**	**53.4**	**52.9**	**53.15**	**51.6**	**45.5**	**48.36**
人工标注 B	58.9	59.1	59.0	62.6	65.9	64.2	50.3	57.69	53.74
人工标注 A	74.3	76.2	75.24	68.5	75.8	71.97	61.3	68.8	64.86

6.2 实体类型聚类对推理的影响

实体类型一致性检测在很大程度上影响事件抽取的性能，而一致性检测的性能取决于两点：实体的聚类性能和相似度计算的准确性。在训练阶段，通过聚类（CLUTO 工具）将同种类型的实体集合按背景知识的异同划分为不同子类；在测试阶段，使用 K 近邻算法，计算实体的背景知识和实体类型描述的相似度，确定实体描述的子类型。

系统从训练语料中获得了 129 种实体子类型。对随机抽取的 10 种子类型测试发现，每种实体子类型都含有超过 19.2%的噪声信息。如表 5 中实体类型“Air”的子类型 1，丢失了实体“*MiGs*”和实体“*enemy*”，同时带来了噪声信息，如实体“*terrorist*”，实体“*Saddam*”等（见表 9）。因此，系统采用人工方法对实体进行聚类并重现跨实体的事件推理过程。图 10 给出了通过人工对实体进行聚类（“Visible 1”）、采用 ACE 中定义的实体类型（“Visible 2”）和采用聚类工具（CLUTO）对实体进行聚类三种方法对应的事件抽取结果（“Blind”）。

表 9 “Air”实体类型的子类型一种的噪音信息（加粗部分）

Fighter plane (Air 实体子类型 1): “*warplanes*” “*allied aircraft*” “*U.S. jets*” “*a-10 tank killer*” “*b-1 bomber*” “*a-10 warthog*” “*f-14 aircraft*” “*apache helicopter*” **“*terrorist*” “*Saddam*” “*Saddam Hussein*” “*Baghdad*”…**

分析表 10 中的数据发现，“Visible 1”对应的实验结果只比“Blind”方法稍好。因此，通过聚类工具对实体进行聚类带来的噪声基本上不会影响事件的推理过程。分析“Visible 1”中的实体子类型发现，

实体子类型的划分粒度仍然不理想，在以后的实验中可进一步细化，如，可以通过进一步加强实体内部信息的一致性（更细粒度的进行实体子类型的划分）改进“Visible 1”中带来的噪音信息。通过对比“Visible 1”和“Visible 2”的结果可以看出，更细粒度的实体类型划分在跨实体推理的事件抽取中更为有效。因此，更有针对性的实体聚类方法将更好的促进跨实体事件抽取的推理过程。

表 10 不同粒度的实体子类型的实验结果

F	触发词识别	事件元素识别	事件角色识别
Blind	68.33	53.15	48.36
Visible 1	69.15	53.65	48.83
Visible 2	51.34	43.40	39.95

7 结论和未来工作

本文针对事件抽取任务提出了一种基于直推式学习的抽取方法，跨实体事件抽取。跨实体事件抽取旨在有效利用实体的一致性特征，进行句子级别的事件触发词识别和事件角色标注。实验结果表明本文的方法明显优于其它基于直推式学习的事件抽取方法，如跨文档事件抽取和跨事件事件抽取。

然而，本文的方法只利用了事件元素的类型信息促进事件角色的标注，并没有有效利用角色之间的上下文信息。比如句子中包含“Attacker”元素，那么事件中会很大概率出现对应的“Target”角色。所以，在以后的工作中可以尝试首先对易于标注的角色信息进行标注，然后利用已标注的角色信息作为先验知识，辅助难以标注的角色信息标注，将会一定程度上提升事件角色标注的性能。

参考文献

[1]. David Ahn. The stages of event extraction[C]. In Proc. COLING/ACL 2006 Workshop on Annotating and Reasoning about Time and Events, Sydney, Australia, 2006: 1-8.

[2]. Ralph Grishman, David Westbrook and Adam Meyers. NYU’s English ACE 2005 System Description[C]. In Proc. ACE 2005 Evaluation Workshop, Gaithersburg, MD, 2005: 5-19.

[3]. Jenny R. Finkel, Trond Grenager and Christopher Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling[C]. In Proc. 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, MI, 2005: 363-370.

[4]. Mstislav Maslennikov and Tat-Seng Chua. A Multi resolution Framework for Information Extraction from Free Text[C]. In Proc. 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, 2007: 592-599.

[5]. Siddharth Patwardhan and Ellen Riloff. Effective Information Extraction with Semantic Affinity Patterns and Relevant Regions[C]. In Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, 2007: 717–727.

[6]. Siddharth Patwardhan and Ellen Riloff. A Unified Model of Phrasal and Sentential Evidence for Information Extraction[C]. In Proc. Conference on Empirical Methods in Natural Language Processing, 2009: 151-160.

[7]. Heng Ji and Ralph Grishman. Refining Event Extraction through Cross-Document Inference[C]. In Proc. ACL:HLT, Columbus, OH, 2008: 254–262.

[8]. David Yarowsky. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods[C]. In Proc. ACL, Cambridge, MA, 1995: 189-196.

[9]. Prashant Gupta and Heng Ji. Predicting Unknown Time Arguments based on Cross-Event Propagation[C]. In Proc. ACL-IJCNLP, 2009: 369-372.

[10]. Shasha Liao and Ralph Grishman. Using Document Level Cross-Event Inference to Improve Event Extraction[C]. In Proc. ACL, Uppsala, Sweden, 2010: 789-797.