**文章编号**：1003-0077(2019)09-0088-08

# 基于联合标注和全局推理的篇章级事件抽取

仲伟峰[1]，杨 航[1,2]，陈玉博[2]，刘 康[2]，赵 军[2]

(1. 哈尔滨理工大学 自动化学院，黑龙江 哈尔滨 150080；
2. 中国科学院 自动化研究所 模式识别国家重点实验室，北京 100190)

**摘 要**：事件抽取可以帮助人们从海量的文本中快速、准确地获取感兴趣的事件知识。然而，目前事件抽取的研究主要集中在从单一句子中抽取事件，由于事件构成的复杂性和语言表述的多样性，多数情况下多句才能完整地描述一个事件。因此，从篇章中抽取出完整的结构化事件信息，显得更有价值和意义。该文首先利用基于注意力机制的序列标注模型联合抽取句子级事件的触发词和实体，与独立进行实体抽取和事件识别相比，联合标注的方法在 $F$ 值上提升了 1 个百分点。然后利用多层感知机判断实体在事件中扮演的角色。最后，在句子级事件抽取的基础上，利用整数线性规划的方法进行全局推理，融合句子级事件信息，实现篇章级事件抽取，与基线模型相比，这种基于全局推理的篇章级事件抽取在 $F$ 值上提升了 3 个百分点。

**关键词**：篇章级事件抽取；联合标注；全局推理

**中图分类号**：TP391 **文献标识码**：A

## Document-level Event Extraction Based on Joint Labeling and Global Reasoning

ZHONG Weifeng[1], YANG Hang[1,2], CHEN Yubo[2], LIU Kang[2], ZHAO Jun[2]

(1. College of Automation, Harbin University of Science and Technology, Harbin, Heilongjiang 150080, China;
2. State Key Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

**Abstract**: Current research on automatic event extraction focuses on sentence-level corpus. However, due to the complexity and the diversity of event description in texts, a complete event is mentioned by multiple sentences in many cases. This paper first proposes an Attention-based Sequence Labeling model for joint extraction of entities and events. Compared with the pipeline of entity extraction plus event recognition, this joint labeling model improves the $F$-score by 1%. Then, we use Multi-Layer Perception to label the entities in the events and identify their roles. Finally, based on the labeling and identification results, this paper leverages integer linear programming for global reasoning, improving the $F$-score of document-level event extraction by 3% compared to the baseline.

**Keywords**: document-level event extraction; joint labeling; global reasoning

## 0 引言

当今社会，互联网已成为大部分人日常生活中不可或缺的一部分，在为人们的生活、学习、工作带来极大方便的同时，互联网中海量的非结构化文本也给用户带来信息冗余繁多的困扰。面对日益增长的非结构化文本数据，如何帮助人们理解并快速获取文本中的知识，显得尤为重要，而信息抽取技术的提出正是为了解决这个问题。作为自然语言处理(Natural Language Processing, NLP)技术中的关键任务，信息抽取在知识获取中扮演着重要的角色。Grishman 等[1]将信息抽取定义为：从自然语言文本中抽取指定类型的实体、关系、事件等事实信息，并形成结构化数据输出的文本处理技术。而面向非结构化文本的事件抽取是信息抽取领域中的关键任务和重要的研究方向(其余还有实体抽取、关系抽取等)，主要应用于事件知识图谱的构建、事件信息获取和辅助其他自然语言理解任务。

事件是个复杂的概念，在不同研究领域有不同的定义。事件抽取领域最具有影响力的评测会议——自动内容抽取(Automatic Content Extrac-

**收稿日期**：2019-01-04 **定稿日期**：2019-04-02

tion，ACE[①])评测会议将事件定义为：事件是发生在某个特定时间或时间段、某个特定地域范围内，由一个或多个角色参与的一个或多个动作构成的事情或状态的改变。事件中的相关术语具体定义如下：

实体(entity)：用户感兴趣的语义对象，通常是一个名词(例如，“人物”)；

事件触发词(event trigger)：触发事件的核心词，通常是动词或者名词(例如，“丧生”或“拍卖”)；

事件元素角色(event argument)：实体在事件中所扮演的角色，即事件的参与者；

事件描述(event mention)：描述事件的一句话或者一个字段，通常会包含触发词和事件元素；

事件类别(event type)：事件触发词和事件角色共同决定了事件的类别。

具体如图 1 所示(数据来源 ACE2005[②])。

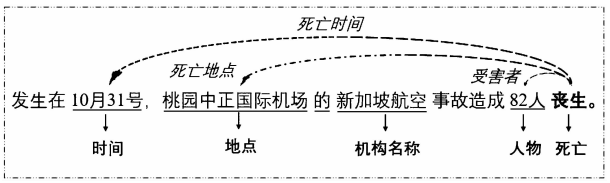


图 1　ACE2005 事件标注实例

如图 1 实例所示，事件触发词和实体描述进行了特殊标记，有下划线的字段代表实体及其类别(例如，“10 月 31 号”，时间)，加粗字段代表触发词及其事件类别(“丧生”，死亡)。虚线连接触发词和实体，其上面文字代表实体在该事件中所扮演的角色。在本实例中，“丧生”触发一个死亡事件，“10 月 31 号”“桃园中正国际机场”“82 人”在该事件中分别扮演时间、地点和受害者的事件角色，从而组成一个完整的事件，而实体“新加坡航空”在该事件中不扮演任何角色。由 ACE 中事件的定义及图 1 实例可得，事件的组成要素主要包括事件的发生时间及地点，事件的参与角色以及与之相关的动作或状态(触发词)。在现实世界中，每天都有各式多样的不同场景、不同类型、不同粒度的事件发生，信息描述多样化的同时也给事件抽取任务带来难度。

作为自然语言处理中具有挑战的任务，事件抽取主要研究如何从非结构化的文本信息中抽取出用户感兴趣的事件，并以结构化的形式呈现出来。目前事件抽取的研究主要集中在两个子任务上：事件识别和事件元素识别。

**事件识别**：识别文本中的由事件触发词引导的事件实例，并根据当前触发词和上下文信息判断当前触发的预定义事件类型。

**事件元素识别**：若某句被判定为特定事件类型的事件描述，需判断句中实体和事件触发词之间的关系，这里的关系即为实体在该事件中所扮演的角色。

上述事件抽取定义主要是针对句子级别的，而现有的事件抽取框架按照文本粒度可分为句子级事件抽取和篇章级事件抽取。句子级事件抽取焦点集中于识别句子中每个词可能提及的单个事件，以及判断句子实体在该事件中扮演的角色。虽然句子级抽取考虑的事件类型足够通用(ACE2005 中定义了 33 种事件)，但对于总结文档内容来说，句子级抽取粒度太细了。现实场景中，一篇文档通常包含一个或者多个事件，这些事件对于整体的重要性各不相同，而同一事件也可能会在文档中被多次提及。篇章级事件抽取以文本中描述的主要事件为中心，用简洁、结构化的形式呈现给用户。其在现实世界中直接面向用户也具有明显的适用性，它允许用户快速获取文档中的事件内容、地点和时间，而不需要通读全文。难点在于，篇章事件抽取需要高质量的句子级抽取结果以及相同事件不同事件描述之间事件元素的融合，考虑以下例句：

**例 1**：根据奥地利救灾组织的统计，在阿尔卑斯山登山缆车失火惨剧中有 155 名乘客丧生。

**例 2**：奥地利一处滑雪胜地的登山缆车 11 号在阿尔卑斯山隧道发生缆车失火惨剧，受害者中包括有 1999 年世界女子花式滑雪冠军施密特。

例 1 和例 2 是描述同一灾难事件的不同句子，分布在原文档中不同的段落当中。例 1 中包含该灾难事件的死亡人数和事故来源，例 2 中包含事件发生的时间和地点。事件描述例 1 和例 2 中的结构化事件信息需要融合才能得到完整的篇章级事件信息，其结构化信息如表 1 所示。

从表 1 可以看出，篇章级事件抽取依赖于句子级抽取结果和跨句子的事件元素融合。从理论出发，为了获取篇章级事件的结构化信息，需要句子级事件抽取结果和事件共指关系判断。目前针对篇章事件抽取研究较少，还没有统一的统计学模型能从篇章中直接抽取出篇章的事件信息。相反，句子级事件抽取的研究日趋成熟，在句子级抽取结果的基础上进行全局推断提高篇章事件抽取的整体性能是

① http://projects.ldc.upenn.edu/ace/

② https://catalog.ldc.upenn.edu/LDC2006T06

**表1 句子级与篇章级事件结构化信息对比**

| 句子 | 句子级事件信息 | 篇章级事件信息 |
|---|---|---|
| 例1 | 事件触发词:丧生<br>受害者:155名乘客<br>事故地点:阿尔卑斯山<br>事故来源:缆车失火 | 事件触发词:丧生、受害<br>受害者:155名乘客、施密特<br>事故日期:11号<br>事故地点:阿尔卑斯山<br>事故来源:缆车失火 |
| 例2 | 事件触发词:受害<br>事故日期:11号<br>受害者:施密特<br>事故地点:奥地利一处滑雪胜地 | |

本文研究的方向。

本文采用管道(Pipeline)的方法将篇章级抽取问题分为3个子问题:①利用序列标注模型对句子进行实体和事件的联合标注;②采用多层感知机对事件描述中的实体进行分类,判断实体在该事件中所扮演的角色;③基于整数线性规划做全局推理,得到篇章级结构化事件信息。在整个流程图中不借助标注语料中的其他信息和外部资源。总的来说,本文的贡献在于以下3点:

(1) 提出了实体和事件的联合标注模型,此模型可以更好地利用上下文中的实体和事件的相互依赖关系。

(2) 提出利用整数线性规划的方法进行全局推理得到篇章事件抽取结果。

(3) 在ACE2005中文语料上进行实验,实验结果验证了模型的有效性。

## 1 方法

近年来,已经证明了神经网络方法在自然语言处理领域的有效性。Zeng等[2],Chen等[3]最先将深度学习的方法应用于关系抽取和事件抽取中,并取得了很好的效果。相对于传统特征表示的方式,神经网络将词向量(Word embedding)作为输入,避免了传统特征提取过程过分依赖词性标注、句法分析等自然语言处理工具。在本节中,我们将介绍本文篇章级事件抽取采取的方法,主要包括实体和事件联合标注、事件元素识别、全局推理。篇章级事件抽取模型整体框架如图2所示。

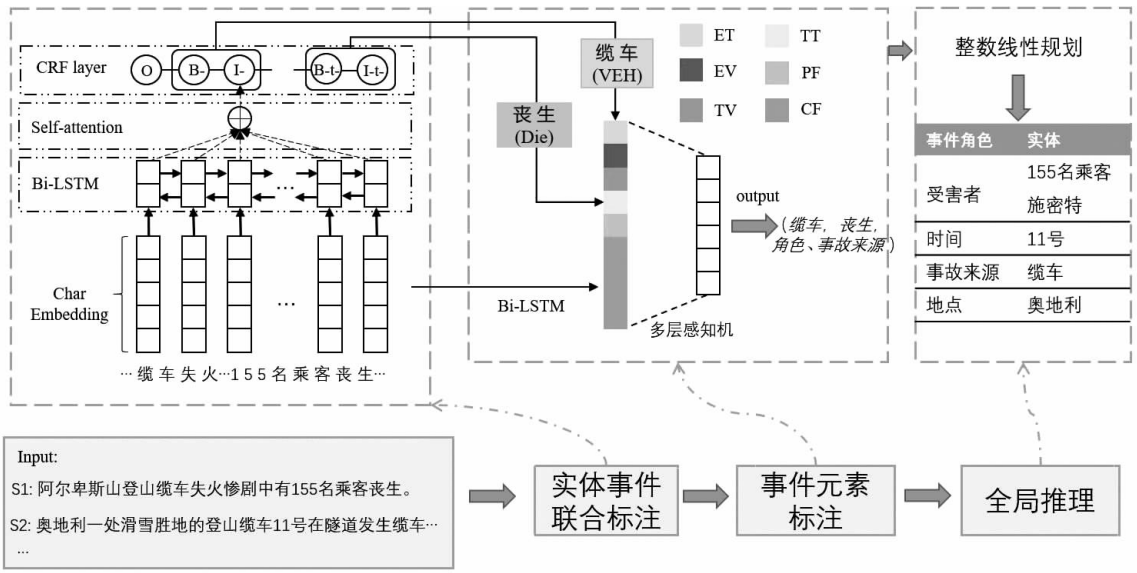


图2 篇章级事件抽取模型整体框架

### 1.1 实体和事件联合标注

如图1描述,实体和事件是紧密关联的,两者的表示相互依赖,但现有的事件抽取通常都对实体和事件分别建模。在目前事件抽取任务中,研究者一般将事件抽取分为3步:①实体识别:利用外部工具或者单一模型抽取句中的实体;②事件识别:抽取句中的触发词并判断事件类型;③元素分类:判断实体在事件中扮演的角色。实体识别和事件识别分开处理是常用的技术手段,但会忽略实体和事件触发词之间相互依赖的关系。例如,在例句"奥巴马离开白宫,迎接新的挑战"中,"离开"作为触发词,触发了一个离职类别的事件而不是运输类别的事件。只分析"离开"一词,会存在歧义,但在已知句中其他实体及其类别后("白宫,组织机构";"奥巴马,人名"),更易判断"离开"触发离职事件;相反,当已知

“离开”触发离职事件，更易判断“白宫”的实体类别是组织机构而不是地理位置。如何有效利用实体和事件触发词的依存关系，是本文提出联合标注模型的出发点。本文采用序列标注模型从句中联合标注实体和触发词，同时判断它们的类别，并将其抽取结果作为事件元素标注的输入。为了更好地建模上下文的关联关系，我们将自注意力机制（self-attention）[4]加到模型中。

目前有很多统计学习方法可以对中文文本中的词进行识别并分类，但利用词作为输入需要先借助外部分词工具，而序列标注方法能够很好地解决中文词间无间隔的问题。在自然语言处理中，很多基础问题都可以用序列标注模型解决，比如中文分词、词性标注以及命名实体识别等。序列标注不仅能捕获词的边界，同时也可以判断当前词的归属类别。不同于文本分类，序列标注模型将输入的句子看作一个序列，输出是一个等长的符号序列，每个符号对应特定的含义。具体来讲，序列标注模型给句子中的每个字符打上 BIO 的标签，B 表示字段开始（beginning），I 表示字段中间（inside），O 表示其他字段（outside），标签后面跟的 type 表示字段的分类结果，例如，B-PER 表示人名的起始字符，I-Attack 表示触发攻击事件词的中间字段。随着深度学习在自然语言处理中的应用日趋成熟，利用神经网络的方法表示字符特征，能更好地捕获字以及上下文的信息。在神经网络中，目前主流的两个方法是循环神经网络（Recurrent Neural Networks，RNN）和卷积神经网络（Conventional Neural Networks，CNN）。相比之下，RNN 比 CNN 更适合给序列进行建模，因为 RNN 的隐层既有当前时刻的输入，也有前一时刻的隐层输出，这使得它能通过循环反馈连接看到前面的信息，并且还具备非线性的拟合能力，因此利用 RNN 对序列到序列的建模是 NLP 中常用的手段。而长短期记忆网络（Long Short-Term Memory，LSTM）能将过去和将来的序列考虑进来，使得上下文信息充分被利用[5]。在 LSTM 后加入条件随机场（Conditional Random Fields，CRF）能更多地考虑整个句子的局部特征的线性加权组合，计算联合概率，优化了整个序列。同时，我们将自注意力机制加到模型中，主要目的是学习句子内部字符之间的依赖关系，捕获句子的内部结构和语义信息，模型如图 2 所示。

### 1.2　事件元素识别

文档中每个句子经过上述的实体和事件联合标注后，可获得句中的实体及其实体类型和事件触发词及其事件类型。为得到句子级的事件结构化信息，需要进一步标注实体在事件中扮演的角色，即实体和触发词之间的关系（例如，判别实体“155 名乘客”在“死亡”事件类型中扮演了“受害者”的角色）。为了充分利用实体特征和句子中的事件信息，本文利用一个多层感知机实现实体的分类从而实现事件元素识别。输入特征包括触发词、触发词类别、实体、实体类别、实体和触发词之间的位置信息以及当前句子通过 LSTM 的向量化表示。

### 1.3　全局推理

在文档文本中，重要的事件通常会被多次提及，即同一事件会有多个事件描述。经过上述句子级事件抽取，可获得篇章中的一系列结构化事件信息。为获得篇章级的事件信息，需要判断多个事件描述是否指代同一事件，从而得到完整的事件信息。如图 2 所示，事件描述例 1 和例 2 分别通过“丧生”和“受害”触发“死亡”事件类型，通过文本描述的相似程度可以进一步判断例 1 和例 2 指代了同一事件，从而将两者的事件元素进行融合得到篇章级的事件结构化信息。为了充分利用文本信息进行事件共指的判断，本文采用整数线性规划的方法进行全局推理，将获取更好的事件共指判断作为优化目标，将文本相似度作为优化目标的重要系数，在条件约束下，得到篇章级事件抽取的最优结果。

## 2　模型

本节主要介绍上述方法所用的模型，包括基于自注意力机制的实体事件联合标注模型、基于感知机的事件元素识别模型和基于整数线性规划的全局推理。

### 2.1　基于自注意力机制的序列标注模型

**输入**：句子中每个字符的字向量表示。输入句子以字符为单位可表示为 $s=\{c_1,c_2,c_3,\cdots,c_i,\cdots,c_n\}$，其中 $n$ 表示该句中字符个数。利用预训练的词向量将句中的字 $c_i$ 映射到低维稠密的字向量 $w_i$，得到句子的向量表示 $s=\{w_1,w_2,w_3,\cdots,w_i,\cdots,w_n\}$ 作为神经网络的输入。

**上下文表示**：Bi-LSTM 生成字符的向量表示，具体来说，LSTM 是由式（1）到式（6）所确定的一类循环神经网络变种。

$$i_t = \sigma(W_{wi} w_t + W_{hi} h_{t-1}) \quad (1)$$

$$f_t = \sigma(W_{wf} w_t + W_{hf} h_{t-1}) \quad (2)$$

$$o_t = \sigma(W_{wo} w_t + W_{ho} h_{t-1}) \quad (3)$$

$$\hat{c}_t = \tanh(W_{wc} w_t + W_{hc} h_{t-1}) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \hat{c}_t \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

双向LSMT将正向和反向的隐层拼接作为下一层输入 $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$。

自注意力机制(self-attention)：学习字符表示之间的依赖关系，捕获句子的内部结构，利用Softmax得到规整化的权重，如式(7)所示。

$$a_t = \text{softmax}(f(Q,K)) = \frac{\exp(Q^{\mathrm{T}} W_a K_t)}{\sum_j \exp(Q^{\mathrm{T}} W_a K_j)} \quad (7)$$

使用得到的权重进行加权求和，得到当前字符的经过Attention后的隐层表示。

$$\text{Attention}(Q,K,V) = \sum_t a_t V_t \quad (8)$$

Self-attention即如式(9)所示，每个字符的表示都会与句中所有的字符进行Attention计算。

$$Q_t = K_t = V_t = h_t \quad (9)$$

**全连接层**：得到网络输出 $P$，如式(10)所示。

$$P = A W_s + b_s \quad (10)$$

**条件随机场**：从式(10)中，得到全连接层的输出为 $P$，其中 $P_{i,j}$ 表示当前字符 $c_i$ 映射到第 $j$ 个标签 $\text{tag}_j$ 的非归一化概率。CRF层中有一个转移概率矩阵 $A$，$A_{ij}$ 表示的是从第 $i$ 个标签 $\text{tag}_i$ 到第 $j$ 个标签 $\text{tag}_j$ 的转移得分，该矩阵的功能是为一个位置进行标注的时候可以利用此前已经标注过的标签。当已知序列 $s = \{c_1, c_2, c_3, \cdots, c_i, \cdots, c_n\}$ 的对应的输出标签结果为 $y = \{y_1, y_2, y_3, \cdots, y_i, \cdots, y_n\}$，我们定义当前序列得分如式(11)所示。

$$\text{score}(x,y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1},y_i} \quad (11)$$

其中，整个序列的打分 $\text{score}(x,y)$ 等于各个位置的打分之和，而每个位置的打分由两部分得到，一部分是由网络层的输出 $P$ 决定，另一部分则由CRF的转移矩阵 $A$ 决定。利用Softmax得到归一化后的概率，如式(12)所示。

$$P(y \mid x) = \frac{\exp(\text{score}(x,y))}{\sum_{y'} \exp(\text{score}(x,y'))} \quad (12)$$

模型通过最大化对数似然函数进行训练，式(13)给出了对一个训练样本 $(x,y)$ 的对数似然。

$$\log P(y \mid x) = \text{score}(x,y) - \log\Big(\sum_{y'} \exp(\text{score}(x,y'))\Big) \quad (13)$$

模型在预测过程(解码)时使用动态规划的Viterbi算法来求解最优路径，如式(14)所示。

$$y^* = \arg\max_{y'} \text{score}(x,y') \quad (14)$$

## 2.2 事件元素识别

利用多层感知机能很好学习各个特征对结构影响的权重参数，是简单高效的神经网络结构。

**输入**：实体，实体类别，事件触发词，触发词类别，文本表示和位置特征。如图2中间部分所示，多层感知机的输入，由各个特征构成，每个特征的具体含义如表2所示。

**表2 分类器输入的特征名称及其含义**

| 特征名称 | 特征含义 |
|---|---|
| 实体(Entity vector, EV) | 实体的向量表示 |
| 实体类别(Entity type,ET) | 实体类别向量表示 |
| 事件触发词(Trigger vector, TV) | 触发词的向量表示 |
| 事件触发词类别(Trigger type, TT) | 触发词类别的向量表示 |
| 文本表示(Context feature, CF) | 当前句子经过LSTM后的隐层表示 |
| 位置特征(Position feature, PF) | 事件触发词和实体的位置距离 |

将上述特征向量化并拼接作为模型的输入，如式(15)所示。

$$X = [\text{EV},\text{ET},\text{TV},\text{TT},\text{CF},\text{PF}] \quad (15)$$

将上述特征作为全连接层的输入，如式(16)所示。

$$O = X W_a + b_a \quad (16)$$

对所有的元素角色类型利用Softmax分类器进行分类，如式(17)所示。

$$p(i \mid x,\theta) = \frac{e^{O_i}}{\sum_{k=1}^{n} e^{O_k}} \quad (17)$$

采用交叉熵作为模型的目标函数，利用梯度下降的方法进行优化，具体目标函数如式(18)所示。

$$J(\theta) = \sum_{i=1}^{T} \log p(y^{(i)} \mid x^{(i)},\theta) \quad (18)$$

## 2.3 全局推理

**输入**：篇章文本中抽取出的句子级事件信息(包含事件文本描述和结构化事件信息)

$$\text{Events} = \{e_1, e_2, e_3, \cdots, e_i, \cdots, e_l\}。$$

**模型**：目标函数如式(19)所示。

$$\text{obj} = \sum_i^l \sum_j^l \text{sim}(e_i, e_j) * \text{var\_refer}_{i,j} \quad (19)$$

式(19)中，$\mathrm{sim}(e_i,e_j)$ 表示 $e_i$ 和 $e_j$ 两个事件的共指程度，依赖事件描述的向量相似度(句子向量化表示后的余弦相似度)和实体的共现程度(句中包含实体的 TF-IDF 余弦相似度)。$\mathrm{sim}(e_i,e_j)$ 为上述两者的线性组合，最终取值[−1,1]。$\mathrm{var\_refer}_{i,j}$ 表示变量。约束函数如式(20)所示。

$$s.t.\ \mathrm{var\_refer}_{i,j} \in \{0,1\} \tag{20}$$
$$\mathrm{var\_refer}_{i,j} = 0 \text{ if } \mathrm{type}_i \neq \mathrm{type}_j$$

约束：若 $e_i$ 和 $e_j$ 的事件类别不同，则它们一定不是同一事件的描述。

在式(20)的约束条件下，最大化式(19)中的目标函数，得到篇章事件结构化信息。

## 3　实验

### 3.1　数据

本文利用 ACE 评测发布的公开语料 ACE2005 中的中文语料作为实验数据集。数据集中标注的实体类别包括：PER(Person，人物)、ORG(Organization，组织机构)、GPE(Geo-Political Entity，政治或人文地理区域)、LOC(Location，地理位置)、FAC(Facility，含有设施的场所)、VEH(Vehicle，运输工具)、WEA(Weapon，武器)以及 VALUE(值)和 TIME(时间)。ACE2005 中预定义 33 个事件子类别，每个事件类别都由不同的事件角色构成。本文参照 Chen 和 Ji 等进行数据的划分[6]，其中 569/64/64/篇文档分别被用作训练集/测试集/验证集。利用 $P$(Precision，精确率)、$R$(Recall，召回率)、$F_1$ 值评价句子级的实体抽取和事件识别性能。参照 Reichart 等[7]采用的篇章级事件抽取评价方式，对于每篇文档，将学习到的结构化事件信息和标准进行最大匹配，然后利用 $P$、$R$ 和 $F_1$ 进行篇章级事件抽取性能的评测。

### 3.2　参数

模型的一些实现细节如下：输入的 embedding 为 100 维的词向量，是通过在维基百科中文语料进行预训练得到的。LSTM 隐层维度为 200，batch 设定为 50，学习率为 0.000 1，droupout 为 0.5，最终采用 Adam 作为优化器。

### 3.3　实验结果

#### 3.3.1　实体和事件联合标注性能

目前中文事件抽取研究较少，且集中在事件识别上。为了证明实体和事件联合抽取的有效性，我们采用相同的模型和基础参数独立进行实体抽取和事件抽取的实验。表 3 给出了基线模型的性能指标和本次实验的评价结果。

**表 3　句子级实体标注和事件识别评价结果**

| 模型 | | $P$ | $R$ | $F$ |
|---|---|---|---|---|
| MEMM | | 78.85 | 48.32 | 59.93 |
| FBRNN | | 57.52 | 42.84 | 49.12 |
| C-BiLSTM | | 60.03 | **60.94** | **60.42** |
| BiLSTM Trigger | | 58.71 | 55.39 | 57.00 |
| BiLSTM Entity | | 71.45 | 77.23 | 74.23 |
| Joint | Entity | 75.50 | 76.39 | 75.94 |
| | Trigger | 60.89 | 55.74 | 58.20 |
| Joint (+Att) | Entity | **77.46** | **78.28** | **77.86** |
| | Tigger | **66.96** | 53.36 | 59.39 |

**MEMM**[6]：运用最大熵马尔可夫模型进行文本序列标注任务；

**FBRNN**：Ghaeini 等[8]提出前后向 RNN 网络进行字符级中文文本事件抽取；

**C-BiLSTM**[9]：该方法结合 CNN 和 LSTM 去捕获文本和词汇信息从而提升中文事件抽取的性能；

**BiLSTM-CRF**：采用双向 LSTM 和 CRF 序列标注的方法进行字符级中文文本事件抽取。

本文提出的实体事件联合标注的方法取得良好的实验效果，并在 $P/R/F_1$ 上都超过了独立进行实体抽取和事件识别模型的性能，验证了该方法的有效性。在加入自注意力机制后，实体标注的 $F_1$ 值提升 1.92%，事件识别的 $F_1$ 值提升 1.19%，该结果验证了利用自注意力机制可增强实体和事件的相互依赖程度，有助于提高实体和事件联合标注的性能。

#### 3.3.2　篇章级事件抽取性能

目前事件抽取的研究大多集中在句子级事件抽取，在篇章事件抽取方面，并没有相关工作在 ACE2005 上进行实验。本文在得到句子级抽取结果的基础上，通过全局推理得到篇章级的结构化信息，选用基线(baseline)如下：

**Base1**：将触发相同事件类别的事件描述看成同一事件；

**Base2**：将事件触发词相同的事件描述看成同一事件。

对于每篇文档，将抽取出的结构化事件信息（事件类别和实体在事件中扮演角色）和标准进行最大化匹配。然后利用 $P$、$R$ 和 $F_1$ 进行篇章级事件抽取性能的评测。表4给出了实验的评价结果。

**表4 篇章级事件抽取评价结果(%)**

| 模型 | $P$ | $R$ | $F$ |
|---|---|---|---|
| Base1 | 28.67 | 29.78 | 29.21 |
| Base2 | 32.32 | 31.84 | 32.56 |
| ILP | **37.68** | **34.47** | **36.00** |

可以看出，基于整数线性规划的模型在篇章级事件抽取结果各项指标上均超过了基线模型。但由于整个模型是基于 Pipeline 的方法，篇章级抽取性能会受到句子级事件抽取结果的影响，造成最终性能指标偏低。

## 4 相关研究

当前事件抽取按照研究方法可分为两大类：基于模式匹配和基于统计模型。模式匹配的方法在特定领域能取到较好的精确度，典型的基于模式匹配的事件抽取系统有：ExDisco[10]和FSA[11]。但该方法需要大量人工进行模板撰写，而且普适性差，只适用于小规模的特定领域。基于统计学习的方法，在特征选取上又可分为两类：基于传统特征选取和基于神经网络自动学习特征。传统特征提取主要通过自然语言处理工具获取各种有效的词汇、句法和语义等特征，然后利用传统分类模型（例如，最大熵模型和支持向量机模型）进行分类[12-14]。随着深度学习证明了其在 NLP 中的有效性，Chen 等[3]率先将 CNN 应用到事件抽取中，并利用了距离信息来建模实体和触发词的位置关系；Nguyen 等[15]提出一种基于 RNN 的模型进行事件识别和角色分类的联合学习。针对语料缺、不平衡等问题，Liu 等[16]借助外部语义资源进行事件识别；Chen 等[17]利用远程监督的方法扩充训练语料提高了事件抽取性能；Yang 等[18]借助篇章信息进行事件和实体的联合抽取，并将其分为3个子问题：学习事件内部结构、学习事件与事件关系、学习实体抽取；Liu 等[19]利用双语资源提高事件抽取的性能。这些方法在英文事件抽取数据集上取得了很好的效果。

中文事件抽取方面，词级的不匹配问题严重影响了汉语信息抽取中词级模型的性能。为了解决该问题，Chen 和 Ji 等[6]提出了基于特征的字符级 BIO 标注；Li 等[20]定义了中文触发词的人工模板，这些方法都高度依赖于人工构建的模板和特征。

从文本粒度看，目前事件抽取的相关研究主要针对句子级别的抽取，即识别句中触发词，并判断实体在事件中所扮演的角色。但现实世界的文本大多是以篇章的形式出现，用户更关心的是从篇章中获得结果化的事件知识。最早的事件抽取系统 FRUMP[21]采用事件模板匹配的方法进行篇章事件抽取。Huang 等[22]采用基于模式分类的方法，将篇章抽取看成两个子问题：①角色槽填充；②句子关联模型。Yang 等[23]采用基于句子抽取结果以及文本特征发现主事件描述，并利用上下文元素补齐策略得到篇章事件结构化信息的方法，在中文金融事件抽取数据集上取得不错的效果。总的来说，目前篇章事件抽取的研究主要集中在特定的领域，高度依赖人工规则，很难推广到新的领域。而句子级事件抽取方法应用于更广泛的领域，但生成的输出粒度太细，无法提供好的文档级事件信息。

## 5 总结和展望

本文讨论了事件抽取对于知识获取的重要性，并阐述了句子级事件抽取和篇章级事件抽取的差异。相比句子级事件抽取的细粒度结果，篇章级事件抽取的结果能反映出完整的事件信息，具有更好的现实意义和实用价值。

为了从文本中获取篇章级事件信息，本文采用深度学习的方法抽取句子级事件信息，其模型由两部分组成：基于序列标注的事件实体联合抽取和基于多层感知机的事件元素识别。在句子级事件抽取基础上，采取整数线性规划进行全局推断得到篇章级事件结构化信息。本文在 ACE2005 数据集上的实验结果证明了方法的有效性。

然而，基于 Pipeline 的方法不可避免地会带来误差的传递。如何利用端到端的模型，从篇章文本中直接抽取出事件结构化信息，提升篇章级事件抽取整体性能，是下一步主要研究方向和内容。

## 参考文献

[1] Grishman R. Information extraction: Techniques and challenges[M]. Information Extraction a Multidisciplinary Approach to An Emerging Information Technolo-

gy. Springer Berlin Heidelberg, 1997: 10-27.

[2] Zeng D, Liu K, Lai S, et al. Relation classification via convolutional deep neural network[C]//Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. 2014: 2335-2344.

[3] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multi-pooling convolutional neural networks [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015, 1: 167-176.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[5] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 2625-2634.

[6] 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. Association for Computational Linguistics, 2009: 209-212.

[7] Reichart R, Barzilay R. Multi event extraction guided by global constraints [C]//Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012: 70-79.

[8] Ghaeini R, Fern X Z, Huang L, et al. Event nugget detection with forward-backward recurrent neural networks [C]//Proceedings of the 54th Computational Linguistics. 2016:369-373.

[9] Lin C Y, Xue N, Zhao D, et al. A Convolution BiLSTM neural network model for Chinese event extraction [C]//Proceedings of Natural Language Understanding and Intelligent Applications Volume 10102. 2016,10: 275-287.

[10] Yangarber R, Grishman R. Customization of information extraction systems [C]//Proceedings of International Workshop on Lexically Driven Information Extraction, 1997: 1-11.

[11] Surdeanu M, Harabagiu S M. Infrastructure for open-domain information extraction [C]//Proceedings of the 2nd International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., 2002: 325-330.

[12] Ahn D. The stages of event extraction[C]//Proceedings of the Workshop on Annotating and Reasoning about Time and Events. Association for Computational Linguistics, 2006: 1-8.

[13] Jungermann F, Morik K. Enhanced services for targeted information retrieval by event extraction and data mining[C]//Proceedings of International Conference on Application of Natural Language to Information Systems. Springer, Berlin Heidelberg, 2008: 335-336.

[14] Liao S, Grishman R. Using document level cross-event inference to improve event extraction[C]//Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 789-797.

[15] Nguyen T H, Cho K, Grishman R. Joint event extraction via recurrent neural networks[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 300-309.

[16] Liu S, Chen Y, He S, et al. Leveraging framenet to improve automatic event detection [C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, 1: 2134-2143.

[17] Chen Y, Liu S, Zhang X, et al. Automatically labeled data generation for large scale event extraction [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, 1: 409-419.

[18] Yang B, Mitchell T. Joint extraction of events and entities within a document context[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 289-299.

[19] Liu J, Chen Y, Liu K, et al. Event detection via gated multilingual attention mechanism [C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[20] Li P, Zhou G. Employing morphological structures andsememes for Chinese event extraction[C]//Proceedings of COLING 2012, 2012: 1619-1634.

[21] DeJong G. Prediction and substantiation: A new approach to natural language processing [J]. Cognitive Science, 1979, 3(3): 251-273.

[22] Huang R, Riloff E. Modeling textual cohesion for event extraction[C]//Proceedings of the AAAI, 2012, 1(2.1): 1.

[23] Yang H, Chen Y, Liu K, et al. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data[C]//Proceedings of ACL 2018, System Demonstrations, 2018: 50-55.

(下转第 106 页)

[20] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the Advances in Neural Information Processing Systems, 2017: 6000-6010.

[21] Vinyals O, Fortunato M, Jaitly N. Pointer networks [C]//Proceedings of the Advances in Neural Information Processing Systems, 2015: 2692-2700.

[22] Kiros R, Zhu Y, Salakhutdinov R R, et al. Skip-thought vectors[C]//Proceedings of the Advances in Neural Information Processing Systems, 2015: 3294-3302.

张家硕(1992—),硕士研究生,主要研究领域为图像描述生成。
E-mail: jiasurezhang@gmail.com

唐建(1993—),硕士研究生,主要研究领域为机器翻译。
E-mail: Johnnytang@gmail.com

洪宇(1978—),通信作者,博士,教授,主要研究领域为信息检索、信息抽取。
E-mail: tianxianer@gmail.com

(上接第 95 页)

仲伟峰(1969—),通信作者,硕士,教授,主要研究领域为模式识别与智能系统。
E-mail: zwfhlg@163.com

陈玉博(1990—),博士,助理研究员,主要研究领域为知识图谱、信息抽取。
E-mail: yubo.chen@nlpr.ia.ac.cn

杨航(1994—),硕士研究生,主要研究领域为自然语言处理、事件抽取。
E-mail: hang.yang@nlpr.ia.ac.cn