中文事件相关性语料库构建及识别方法

黄一龙[1,2]，李培峰[1,2]，朱巧明[1,2]

（1.苏州大学 计算机科学与技术学院，江苏 苏州，215006；

2.江苏省计算机信息处理技术重点实验室，江苏 苏州，215006）

摘要：事件往往围绕主题展开，相互间存在相关性。在大数据时代，从海量信息中筛选出和某个主题相关的事件，有助于信息抽取、文本摘要、文本生成等自然语言处理任务。本文首先提出一种相关事件的标注方法，并标注了一个中文事件相关性语料库。然后，初步提出了一个基于多种特征的相关性事件识别方法。在标注语料上的实验表明，性能在基准系统上 F1 值提高了 4.08%。

关键词：相关事件语料库；标注；相关性；事件关系

中图分类号：TP391　　**文献标识码：**A

The Construction of Chinese Relevant Event Corpus and Its Recognition Approach

HUANG Yilong[1,2] ,LI Peifeng[1,2] ,ZHU Qiaoming[1,2]

(1.School of Computer Science and Technology , Soochow University ,Suzhou ,Jiangsu 215006 ,China ;

2.Province Key Lab of Computer Information Processing Technology of Jiangsu ,Suzhou ,Jiangsu 215006 ,China)

Abstract: There are many events surrounding a topic in a document and they are relevant events. In the era of Big Data, extracting those events which are relevant to a specific topic is helpful for many natural language processing applications, such as Information Extraction, Text Summarization, and Text Generation. Firstly, this paper proposes a method to annotate relevant events and constructs a Chinese relevant event corpus. Secondly, it puts forward a relevant event recognition approach based on various kinds distance and semantics features. The experimental results on the annotated corpus show that our approach outperforms the baseline 4.08% in F1-measure.

Key words: relevant event corpus; annotation; relevance; event relation

1 引言

如今，互联网已成为人们交流信息和获取资料的主要平台。在其为人们生活提供便利的同时，每天还会产生海量数据。这些数据中有大部分以文本形式存储，记录了大量事件，而其中有许多事件相互关联。当人们使用搜索引擎查找某个特定事件时，不但希望获取所关心的事件，还希望能够获得与之相关的事件。

事件是描述特定目标在某个时间、地点的某种状态。ACE[1]对事件作如下定义：事件是包含参与者的具体发生的事情，常被用来描述状态的改变。在事件之间的多种关系中，时序关系、因果关系等方面的研究相对较多，而在相关关系方面的研究较少。舍恩伯格在《大数据时代》中认为，建立在相关关系分析法上面的预测是大数据的核心，只有在完成了对相关

基金项目：国家自然科学基金（61472265）；国家自然科学基金重点项目（61331011）；江苏省前瞻性联合研究项目（BY2014059-08）；软件新技术与产业化协同创新中心部分资助

[1] ACE Guidelines 5.5.1, http://www.ldc.upenn.edu/Projects/ACE/

性分析后，才有可能研究更深层次的因果关系，找出背后的为什么。而且，由于因果关系较为复杂，对它的定义理论学界还存在争议，标注一个因果关系的语料库存在很大困难。为此，本文从事件的相关性入手研究事件关系。

在一篇文章中，事件围绕主题为中心展开，文章内容由其所包含的各个事件进行描述，因此这些围绕一个特定主题的事件在一定程度上具有关联性。其中关联性大的事件称为相关事件。一般情况，一篇文章可看成由几个话题组成，话题中的事件往往是相关的。

*例 1：国民党副主席吴伯雄最近**访问(E1)**了中国大陆，并且**会见(E2)**了中国高层领导人。*

(——VOM20001129.0700.2200)

*例 2：以色列士兵 7 号在以色列和黎巴嫩的边界对一群**扔(E1)**石头的巴勒斯坦示威者**开火(E2)**，结果造成两名巴勒斯坦人**丧生(E3)**和 10 多人**受伤(E4)**。*

(——CBS20001008.1000.0742)

在例 1 中，“访问”和“会见”具有顺序性，并且后者是前者的目的。在例 2 中，“开火”是“丧生”和“受伤”的原因。容易看出，相关事件往往有相同或相似的论元，而且事件类型也较固定。例如，在 ACE2005 中文语料库中，“Start-Position”(任职)类型事件与“Be-Born”(出生)、“Marry”(结婚)等类型的事件无关。

本文提出一种中文事件相关性语料库构建方法，并提出基于多种特征的事件相关性识别方法。正确识别两个事件的相关关系，可以更好地表示文章主题，将事件中的时间、地点、角色、类型等语义信息相关联。这有助于人们获取更多与该事件相关的信息。同时，信息抽取、文本摘要、文档生成、自动问答等任务也能够根据其相关事件，提取到更多有价值信息。

本文组织结构如下：第 2 节介绍相关工作；第 3 节介绍相关性语料库构建及标注结果；第 4 节介绍针对识别事件相关性提出的特征；第 5 节为事件相关性识别的实验结果；最后一节对本文工作进行总结，并展望将来的工作。

2 相关工作

马彬[1]对相同话题收集多篇文章，使用依存关系构建依存线索集。根据线索集计算事件依存强度，以判断标题事件相关性。杨雪蓉[2]在马彬的语料上，提出利用核心词、依存实体、共现实体关联因子，构造事件关联因子，通过事件关联因子的大小判断标题事件对是否相关。

Zou[3]提出一种中文事件模式标注方法，他们定义了 7 种事件关系，分别为：因果关系(Causality)、同指关系(Co-reference)、顺序关系(Sequential)、目的关系(Purpose)、部分-整体关系(part-while)、并列关系(Juxtaposition)、对比关系(Contrast)。并且认为同一篇报道中的事件是有一定关系的。

Chanmbers[4]提取了事件对之间的特征，如词法特征、句法特征等，使用 SVM 分类器进行事件对时序关系的识别。Chanmbers[5]使用 ILP(Integer Linear Programming)方法，对识别结果进行全局优化，将互相矛盾的结果进行重新识别。Ittoo[6]对语料进行句法分析后，利用少量已知的因果模板，寻找新的因果关系，再利用新找到的因果关系继续寻找新的模板，在语料库中循环迭代用于抽取因果关系和模板。Sorgente[7]使用模板匹配的方法找出所有可能含有因果关系的句子，使用规则方法抽取这些句子中可能的原因和结果，最后使用词汇、语义、依存特征，对是否有因果关系进行分类。

Wolff[8]使用条件概率表示两个事件之间的因果关系强度，并给出 4 种事件关系：因果(Cause)，促进(Enable)，阻止(Prevent)，抑制(Despite)。给出判断这 4 种关系的 3 个特征：受影响因素(patient)是否有到达目标状态的趋向，影响因素(Affector)与受影响因素的出现或缺失是否一致，是否达到目标状态。

相关工作中，事件关系研究集中在因果关系与时序关系，而中文事件关系识别较少。在事件相关关系方面，马彬[1]、杨雪蓉[2]侧重使用无监督方法，利用文档内容信息，识别相同话题下，不同文档标题事件之间的相关性。本文侧重标注和识别同一文档内事件相关性，而同一文档内可能含有多个话题。

3 相关性语料库构建

因果关系局限于事件类型，且在定义上分歧较大，本文从相关性角度研究事件关系。相关性概念广泛，可认为万物相关，也可认为万物不相关。因此，需要对相关性制定准则。本节提出基于子话题的事件相关性标注规则，将细粒度的传统事件关系进行粗粒度标注。目的是为了保证标注一致性和下一步的识别工作。

3.1 语料库标注规则

传统的事件关系可以分为 7 类[3]：因果关系、同指关系、顺序关系、目的关系、部分-整体关系、并列关系、对比关系。

本文将上述关系简化，将属于上述 7 种关系的事件对标注为相关，否则标注为不相关。

子话题：文章通常围绕某一主题，从多个方面进行叙述。假设文章中若干事件围绕某一事件展开，描写该事件的过程、后续、结果等信息，则将该事件作为子话题，而描写子话题的若干事件作为其内容，使文章在局部形成层次结构。

在标注时，采用如下规则进行标注：

1) 将文章分为多个子话题。

2) 对每一个事件，判断其属于哪一个子话题，将同一子话题的事件归类。

3) 对于不同子话题下的事件，标注为不相关；对于相同子话题下的事件，根据其事件的触发词、事件类型、论元等信息，判断是否属于 7 类事件关系中的一种，如果是，则标注为相关，否则标注为不相关。

例 3：

*他同时称，阿拉法特还将同克林顿讨论以色列对巴勒斯坦人民的持续**侵犯(E1)**的问题。……另据报道，当天在加沙地带和约 旦河西岸地区仍有零星的**冲突(E2)**发生，已经造成了 2 人**死亡(E3)**，10 多人**受伤(E4)**。……在西伯伦市中心的犹太人定居点外，巴勒斯坦示威者与警察发生了**冲突(E5)**。* (——CTV20001106.1330.1311)

分析文章内容，“冲突(E2)”和“冲突(E5)”由“侵犯(E1)”引出，而“死亡(E3)”和“受伤(E4)”由“冲突(E2)”引出。因此将 E1，E2，E5 作为子话题，而 E3，E4 作为 E2 的内容。可以将例 3 中的事件层次结构化，如图 1 所示：

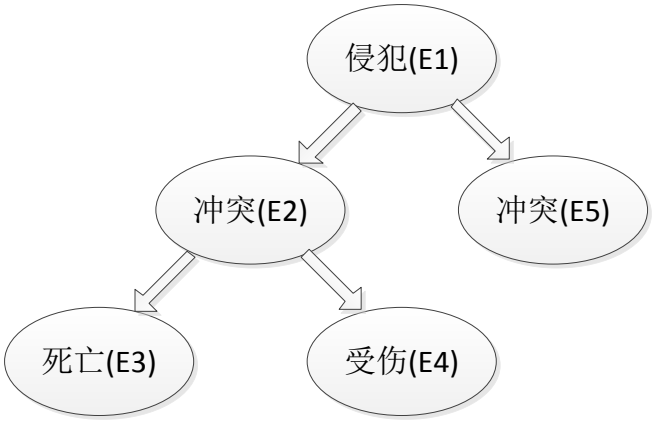

图 1：文章局部层次结构

从图 1 可看出，同一子话题下的事件在树型结构上为兄弟、祖先、后代关系，只有满足上述关系时，才可能相关，否则不相关。

根据标注规则，(E1，E2)、(E1，E5)为部分-整体关系，标注为相关。(E2，E3)、(E2，E4)为因果关系，标注为相关。(E3，E4)为并列关系，标注为相关。(E1，E3)、(E1，E4)可通过 E2 做媒介，构成因果关系，标注为相关。而(E3，E5)、(E4，E5)在不同子话题下，标注为不相关。

对于 ACE2005 中定义的 Transport(移动)类型，由于这类事件可以出现在任何主题的文档中，且表述形式多样。

*例 4：警方据报赶到了现场，将头部被殴打成**伤(E1)**的林姓男子**送(E2)**医急救。*

(——CTS20001206.1300.0398)

*例 5：这 9 名逃犯是 22 号从曼谷西南 40 公里的一所监狱**越狱(E1)**的。*

(——CTV20001123.1330.1541)

*例 6：哥伦比亚武装分子日前**越过(E1)**边境，在巴拿马大林省杀害一名 12 岁儿童并**打(E2)伤(E3)**12 人。* (——CBS20001016.0800.0768)

上述例子中，“送”、“越狱”、“越过”事件在语料中为 Transport 类型事件，但是触发词意义却相差甚远。这类事件对标注工作带来很大干扰，因此本文规定，Transport 类型的事件与另一事件为目的关系或因果关系时，才将其关系标注为相关。如在例 4 中，(E1，E2)存在因果关系，因此将其标注为相关。例 6 中，“越过”只是发生在“打”、“伤”之前，即顺序关系，并无必然联系，因此将(E1，E2)和(E1，E3)标注为不相关。

在标注时有以下性质：

1) (E1，E2)相关$\Rightarrow$(E2，E1)相关

2) (E1，E2)不相关$\Rightarrow$(E2，E1)不相关

3) (E1，E2)相关，(E2，E3)相关$\not\Rightarrow$(E1，E3)相关

性质 1)和性质 2)说明相关关系具有对称性，这是显然的。而性质 3)说明相关关系没有传递性，因为如果使用传递性，结合对称性，会将大量没有关系的事件对标注为相关。在例 3 中，(E1，E3)相关，(E1，E5)相关，但是(E3，E5)不相关。

3.2 语料库标注结果

本文选取 ACE2005 中文语料库作为基础，ACE2005 中文语料库中的语料有三个来源，分别为 broadcast news、newswire、weblog。本文选取来源为 broadcast news 的文档进行标注，因为该类文档中每个文档往往存在多个不同的话题。broadcast news 文档共 298 篇，包括 1398 个事件实例。本文对每篇文档的所有事件进行两两组合，剔除互为同指关系的事件对(互为同指关系的事件必为相关事件)，共构成 9300 个事件对。本文对这 9300 个事件对进行标注。

标注过程由两位标注者分别独立完成，一位是标注规则制定者，在标注前已对所要标注的语料已有一定了解和研究。另一位标注者主要研究事件时序关系，对 ACE 语料库的内容比较了解，但是对事件相关关系没有研究。最终标注结果如表 1 所示：

表 1：语料库标注结果

事件关系	相关	不相关
标注个数	4940	4360
比例	53.12%	46.88%

本文使用 Kappa 值作为衡量标注一致性指标，最终 Kappa 值为 78.18%。分析标注不一致的事件对，发现判定子话题存在较大歧义。

*例 7：**冲突(E1)**已经使 100 多人丧生，其中大部分都是巴勒斯坦人。在星期三的几次**会谈(E2)**上……。* (——VOM20001018.1800.0119)

在例 7 中，规则制定者认为“冲突”为子话题且导致“会谈”事件，因此(E1，E2)相关。而另一位标注者认为(E1，E2)不相关。因此，确定子话题是标注工作中非常重要的环节。

4 事件相关性识别

本节在第 3 节的基础上，参考事件时序、因果等关系的工作，提出基本特征构建基准系统。并提出一系列扩展特征，可分为位置、词汇、句子、类型特征。

4.1 基本特征

本文参考事件时序、因果关系[5][7][8][10]的工作，提出以下特征构建基准系统。

1) 事件特征：事件触发词、类型、子类型、形态、极性、泛型、时态

2) 论元特征：事件对是否有相同论元

3) 句子级特征：事件对的句子距离，事件对的句法路径、依存路径

4.2 扩展特征

在基准系统上，本文提出 7 种扩展特征。可分为位置、词汇、句子、类型特征。

*例 8：这 3 名反对党领导人这次**被捕(E1-1)**前保释在外，有关方面下令让他们明年 4 月出庭**受审(E2-1)**。反对党领导人指责马哈蒂尔政府因为在议会选举中失利而下令**逮捕(E1-2)**这 3 个人。* (——VOM20001223.0700.0222)

以下内容多次涉及同指事件这一概念。同指事件是对一个事件实例的多次描述。在例 8 中，被捕(E1-1)与逮捕(E1-2)指的是同一个事件，因此它们互为同指事件。

1) **位置特征：**

事件对最短距离：事件对能够通过同指事件达到的最短距离。统计语料时发现当两个事件在同一句话内时，相关的概率最大，达到 68.89%，其次是两个事件分布在相邻句子中。因此考虑通过同指事件缩短事件对距离。在例 8 中，E1-2 与 E2-1 的句子距离为 1，但是 E1-1 与 E2-1 的句子距离为 0，小于 E1-2 与 E2-1 的距离，因此该特征记为 0

包含特征：是否存在一个事件，它在文章中出现数次(及存在同指事件)，且能够在位置上包含待识别的两个事件。由于本文语料库多数来自新闻报道，而新闻的特点是将文章的子话题分为“总-分-总”的形式，即先提及子话题，再描述其细节，最后做总结，因此子话题常常包含其相关事件。在例 8 中，事件出现顺序为 E1-1，E2-1，E1-2，可以看到 E2-1 被 E1-1 和 E1-2 包含，因此该特征为 true。

2) 词汇特征

触发词相似度：使用 HowNet[2]计算两个事件触发词的相似度。

连接词：两个触发词之间是否有连接词(词性为 CC 或 CS，“造成”、“结果”等)，并且只判断同一句子内或相邻句子内是否有连接词。

3) 句子特征

句子相似度：对事件对所在句子进行分词，事件 1 所在句子分词后单词数为 num1，事件 2 所在句子单词数为 num2，它们有 num 个相同单词，设

[2] http://www.keenage.com/html/e_index.html

$$sim(num1, num2, num) = \frac{num}{\sqrt{num1 \cdot num2}} \quad (1)$$

当 $sim(num1, num2, num) > \alpha$ 时，将特征置为 high，否则置为 low。

句子相似度衡量两个句子分词后相同词的数量，事件涉及的实体可能会在句子中多次提到，因此这些多次出现的实体在计算相似度时会被多次计算，放大实体的影响。

4) 类型特征

类型是否相同：两个事件类型是否相同，如果相同，则置该特征为事件类型的值，否则将值置为 false。在例 8 中，E1-1 与 E2-1 的类型都为“Justice(审判)”，因此该特征为 Justice。

子类型是否相同：两个事件子类型是否相同，如果相同，则置该特征为子类型的值，否则将值置为 false。很显然，类型或子类型相同的事件往往是一篇报道中的子话题(泛型属性为 Generic)与具体事件(泛型属性为 Specific)。在例 8 中，E1-1 与 E2-1 的子类型不同，分别为“Arrest-Jail(逮捕)”、“Trial-Hearing(审理)”，因此该特征为 false。

5 实验及结果

5.1 实验设置

本文使用 ICTCLAS2015[3]进行分词，Stanford Parser[4]进行句法分析和依存分析。使用 Mallet[5]工具包中的最大熵分类器，按文档进行 5 倍交叉验证。每次实验取 1/5 文档作为开发集。使用正确率(Accuracy)、准确率(Precision)、召回率(Recall)、F1 值作为系统性能评价指标。使用的特征中，事件基本特征、同指事件、论元在 ACE2005 中文语料库已标注，可以直接从语料库中抽取使用。

5.2 实验结果及分析

经开发集调试，句子相似度特征取 $\alpha = 0.2$。实验结果如表 2 所示。从实验结果中可看到，扩展特征能提升系统识别性能。

表 2：事件相关性识别结果

	Accuracy(%)	Precision(%)	Recall(%)	F1
基准系统	68.54	69.34	73.08	71.16
基准系统+事件对最短距离	69.06	69.60	74.15	71.80(+0.64)
基准系统+包含特征	68.96	69.83	73.18	71.46(+0.3)
基准系统+触发词相似度	70.03	70.68	74.47	72.53(+1.37)
基准系统+连接词	68.82	69.89	73.16	71.49(+0.33)
基准系统+句子相似度	70.16	70.99	74.11	72.52(+1.36)
基准系统+类型是否相同	70.95	72.09	73.93	73.00(+1.84)
基准系统+子类型是否相同	71.22	71.86	75.30	73.54(+2.38)
基准系统+以上全部特征	73.08	73.54	77.02	75.24(+4.08)

经验证，使用不同特征组合进行实验，基准系统+所有特征的识别性能为最优。

[3] http://ictclas.nlpir.org/downloads
[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] http://mallet.cs.umass.edu/

从实验结果可以看出：

1) 加入事件对最短距离特征能使性能提高 0.64%，表明跨句子识别性能较低，通过缩短事件对距离能够提高识别性能。

2) 句子相似度特征能明显地提高识别性能，达到 1.36%。基准系统使用论元相同特征，但是句子中的实体未必是事件的论元。从性能提升中可发现，该特征的确能够捕获句子中的实体信息。

3) 包含特征和连接词特征分别能使性能提高 0.3%和 0.33%，提升并不明显。分析识别错误的事件关系发现，虽然包含特征能很好地识别相关性，但是只占总数的 49.57%。而连接词特征噪音较大，对于长句子，连接词可能并不连接事件对。

4) 类型和子类型是否相同的加入，能使识别性能有明显提升，分别达到 1.84%和 2.38%。当两个事件对的类型(子类型)相同时，并不给出一个 true 标记，而是给出具体类型(子类型)的值，目的是捕获 Movement(Transport)类型事件，因为该类事件大多不相关。

5) 所有特征加入后，系统性能提升 4.08%，可以看出，这些特征有效。但是性能提升总量明显少于所有单个特征提升的总和，原因是这些特征并不独立，它们之间相互重叠。如触发词相似度，类型、子类型是否相同特征的重叠性大，因为触发词相似度大的事件，往往有相同的类型和子类型。而句子相似度与基准系统中的事件距离特征有较大重叠，因为如果两个事件原先就在同一句子内，那么句子相似度必定非常高，而跨句子且相似度高的事件对仅占 19.24%，这大大影响了跨句子的识别性能。

6 总结

本文首先提出了一种构建相关性语料库的方法，并且从 ACE2005 中文语料库中选取 298 篇，经过标注后形成本文实验所用语料。其次在标注结果上，提取事件对之间的特征，使用分类器进行相关性识别，使性能 F1 值提高 4.08%。

下一步工作可以从以下方面对本文工作进行扩展。

1) 对无法缩短距离的事件对，抽取更多句子级别特征，提高跨句子事件对的识别性能。

2) 将具有相关关系的事件对进行细粒度划分，从而更好地刻画事件关系。

3) 考虑使用图模型，以引入文档中其他事件信息，充分利用句子中的实体，避免孤立地看待两个事件对之间的特征。

参考文献

[1] 马彬, 洪宇, 杨雪蓉, 等. 基于语义依存线索的事件关系识别方法研究[J]. 北京大学学报: 自然科学版, 2013, 49(1): 109-116.

[2] 杨雪蓉, 洪宇, 马彬, 等. 基于核心词和实体推理的事件关系识别方法[J]. 中文信息学报, 2014, 28(2): 100-108.

[3] Zou H, Yang E, Gao Y, et al. The Annotation of Event Schema in Chinese[C]//23rd International Conference on Computational Linguistics. 2010: 72-79.

[4] Mirza P, Sprugnoli R, Tonelli S, et al. Annotating causality in the tempeval-3 corpus[C] //Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL). 2014: 10-19.

[5] Chambers N, Wang S, Jurafsky D. Classifying temporal relations between events [C]//Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, 2007: 173-176.

[6] Chambers N, Jurafsky D. Jointly combining implicit constraints improves temporal ordering[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2008: 698-706.
[7] Ittoo A, Bouma G. Extracting explicit and implicit causal relations from sparse, domain-specific texts[M]//Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2011: 52-63.
[8] Sorgente A, Vettigli G, Mele F. Automatic Extraction of Cause-Effect Relations in Natural Language Text[J]. DART@ AI* IA, 2013, 2013: 37-48.
[9] Wolff P. Representing causation[J]. Journal of experimental psychology: General, 2007, 136(1): 82-88.
[10] Bethard S, Corvey W, Klingenstein S, et al. Building a corpus of temporal-causal structure[J]. Sixth International Conference on Language Resources & Evaluation Lrec, 2008, 24(1):908-915.
[11] Rink B, Bejan C A, Harabagiu S. Learning Textual Graph Patterns to Detect Causal Event Relations[J]. 2010.
[12] 仲兆满, 刘宗田, 周文, 等. 事件关系表示模型[J]. 中文信息学报, 2009, 23(6): 56-60.

黄一龙（1991—），男，硕士研究生，主要研究方向为中文信息处理。
Email:yilonghuang123@163.com

李培峰（1971—），男，教授，主要研究方向为中文信息处理。
Email:pfli@suda.edu.cn

朱巧明（1963—），男，教授，主要研究方向为中文信息处理。
Email:qmzhu@suda.edu.cn