

## 基于语义事件因果关系识别

杨竣辉<sup>1,2</sup>, 刘宗田<sup>1</sup>, 刘 炜<sup>1</sup>, 苏小英<sup>1</sup><sup>1</sup>(上海大学 计算机工程与科学学院, 上海 200072)<sup>2</sup>(江西理工大学 信息工程学院, 江西 赣州 341000)

E-mail: jwcjhy@126.com

**摘 要:** 文本中因果关系是常见而重要的一类关系, 传统的因果关系识别主要依托因果关系连接词或大量的因果关系特征向量, 该方法适应性不强. 针对这一问题, 提出一种新的方法识别文本中的因果关系, 先文本以事件划分, 构建事件及事件要素间的语义关联性. 抽取具有相关性事件对作为判定因果关系的候选事件对. 再根据因果关系存在时间优先性、因果概率评价性、因果关系依赖性等特征构造计算等式, 并结合事件要素的语义信息计算候选事件间的因果关联度(RCE), 并根据因果关联度设定阈值作为判定事件对是否是因果关系的二元分类器. 实验表明, 该方法对识别因果关系事件对效果较好.

**关键词:** 事件; 因果关系; 因果特征; 因果关联度

**中图分类号:** TP391

**文献标识码:** A

**文章编号:** 1000-1220(2016)03-0433-05

## Identify Causality Relationships Based on Semantic Event

YANG Jun-hui<sup>1,2</sup>, LIU Zong-tian<sup>1</sup>, LIU Wei<sup>1</sup>, SU Xiao-ying<sup>1</sup><sup>1</sup>(School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)<sup>2</sup>(School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China)

**Abstract:** The causal relationship is a kind of common and important relationship between the texts, traditional causality realtions identification relies on the causal relationship connectives or causal features vector, and this method is short of adaptability. In order to deal with this problem, this study proposees a new method to identification causal relationship. First, Partition text use event can constructed event and event factors semantic relevance between events and extracting relevant events pair as a candidate events pair for determinant causal. Then according to causal relationship property of time priority, causal probabilistic evaluation, causal dependence etc. construction calculation formulation and combine element semantic information of the events to calculate the candidate event pair causal correlation degree (RCE), and according RCE set threshold as a binary prediction to prediction an event pair as causal or not. Experiments show that, this method is good for identify the causal relationship events pair.

**Key words:** event; causality; causal features; RCE

## 1 引言

事件因果关系作为文本中既常见而又非常重要的一类语义关系, 有着广泛的应用前景. 因果关系是反映事件间的一种关系. 它可以帮助了解事件发生所涉及因素, 获取事件演变的过程. 因果关系的识别有助于对事件发生的进一步认识, 从而有助于事件后期发展的预判, 为决策者提供重要的信息. 因果关系可分为显式因果关系与隐式因果关系, 这两种表现形式常出现在不同背景下的文本中, 区别它们的关键是语句间是否存在“因为, 由于, 造成”等因果连接词. 如“这场车祸造成3人死亡, 5人受伤”, 对于显式因果关系大多数可直接通过连接词直接映射事件间的关系; 然而诸如“在这场火灾中, 3名消防队员受伤”等例子, 语句间没有显式连接词的因果关系事件, 则需根据上下文内容或语义特征, 来判断事件间是否存在因果关系, 这无疑加大其识别难度. 为此本文提出利用事件及事件要素间的语义关联性来判定文本中隐式因果关系事

件, 从而有助于更全面的理解事件关系.

## 2 相关工作

事件因果关系的研究由来已久, 早在 Aristotle 的《Metaphysics》一书中就有过详细阐述<sup>[1]</sup>. 此后, 科学家在哲学与逻辑学上均围绕着这个问题进行了更深入的探讨, 但直到近代, 因果关系才在计算机领域中有了定性和定量的研究.

Garcia<sup>[2]</sup>通过分析法语中表达因果关系的动词, 实现了一个 COATIS 系统用于抽取法语中形如(Cause Verb Effect)的显式因果关系. Khoo 等<sup>[3]</sup>认为除了动词之外, 还有些连词、形容词、副词以及子句也能表达因果关系, 并采用模式匹配的方法从手工标注的 Wall Street Journal 语料中抽取显式因果关系. Blanco 等<sup>[4]</sup>则采用了模式匹配和机器学习相结合的方法从 TREC5 语料上抽取显式因果关系. 干红华等<sup>[5]</sup>提出了一种事件因果关系结构分析方法, 并且采用产生式规则实现了

收稿日期: 2014-12-10 收修改稿日期: 2015-05-25 基金项目: 国家自然科学基金项目(61273328, 61305053, 31260292)资助; 上海市自然科学基金项目(12ZR1410900)资助. 作者简介: 杨竣辉, 男, 1981年生, 博士研究生, 副教授, 研究方向为知识表示、自然语言处理和 Web 数据挖掘等; 刘宗田, 男, 1946年生, 博士生导师, 研究方向为人工智能和软件工程等; 刘 炜, 男, 1978年生, 博士, 副研究员, 研究方向为语义本体、知识表示等; 苏小英, 女, 1980年生, 博士研究生, 讲师, 研究方向为数据挖掘、自然语言处理等.

一个辅助法律咨询系统. Girju 等<sup>[6]</sup>通过 Internet 和 WordNet 搜寻因果关系动词,建立 Lexico-syntactic 模式,实现了特定事件因果关系的自动识别,且部分解决了运用 C4.5 决策树在语法和语义限制下,学习带标记语料问题.

在这些显式因果关系事件的抽取中,通常采用模式匹配的方法将文本中的事件因果关系抽象为某些特定的模式,这些方法都有很强的领域性,需要大量人工标记.为使得研究成果适应更广泛,近年来有研究者开始尝试着利用统计概率的方法从文本中抽取显式因果关系.

Marcu 等<sup>[7]</sup>采用朴素贝叶斯(Naive Bayes)方式,在分析相邻句子间的词对概率情况下去提取因果关系. Inui 等<sup>[8]</sup>通过因果关系话语指示词(Indicator)和句子之间语言成分,通过支持向量机(Support vector machine, SVM)去识别语料中显式的因果关系.钟军等<sup>[9]</sup>提出基于双层模型(Cascaded model)方法对突发事件显式因果关系抽取,有效地解决了因果关系标注问题和对应关系问题,获得较好的效果.

虽然上述方法的效果较好,识别率基本上均超过 85%,但上述方法也仅限于抽取带有标记的(形如“因为”、“由于”、“造成”等)句内的或相邻句子间的因果关系.事实上文本的描述过程还存在大量隐式的因果关系.为此,也有研究者着手研究抽取语料中隐式因果关系.

[B. Steven]等<sup>[10]</sup>开始尝试用机器学习的方法从手工标注的 TimeBank 语料中抽取显式和隐式标记的因果关系.[Sun]等<sup>[11]</sup>提出通过搜索时间日志对来发现因果关系.郑皎凌[等]<sup>[12]</sup>提出基于扰动的亚复杂动力系统因果分析方法挖掘因果关系.[Beamer]等<sup>[13]</sup>根据电影剧本语料库中的动词设法发现因果关系,但局限于连贯的动词或相邻的词组之间.[Riaz]等<sup>[14]</sup>提出基于“控制依赖效应”(EDC)度量方法,识别事件间的显式与隐式因果关系.Quang 等<sup>[15]</sup>提出基于集中分布的相似性的方法和篇章联系语的语境(CEA)实现事件之间的显式与隐式因果关系识别.付剑锋等<sup>[16]</sup>提出利用层叠条件随机场(CRFs)去识别事件间的显式与隐式因果关系.

综合上述研究,采用单一的模式识别方法进行因果事件抽取的适应性不强,基础局限于特定领域,同时这种方法需要大量人工标记作保证,需要时间代价和经济代价,而采用机器学习方式则需要大量标记训练语料,且该方法仅对带显式连接词标记的因果关系抽取效果较好.在隐式因果关系的研究中,多数研究者在事先对因果关系标注的基础上进行识别,采用基于统计的方式识别句间或跨句文本间的因果事件,尽管效果较好,但过分依赖特征向量,导致该方法的通用性不强.

本文在他们所做的工作中得到启发,提出基于事件语义的因果关系抽取方法.主要针对抽取显式或隐性的因果关系作为研究对象.将文本按事件划分,鉴定事件及事件的角色,接着通过计算事件的语义角色(六要素)间关联度来判定事件间的关联度;接着根据事件间因果关系存在时间优先性、因果概率评价性、因果关系依赖性等特征分别列举相应的计算等式计算候选事件间的因果关联度并设定因果关系的阈值作为判定事件间因果关系的二元分类器.实验表明,这种方法能有效的抽取出文本中的因果关系事件对,不限句子间是否存在因果关系连接词,同时该方法对事件领域特征要求不高,既适用于句子间的因果关系抽取,对跨句的因果关系抽取仍适用.

### 3 事件的相关定义及抽取

#### 3.1 事件及事件关系定义

定义 1. 事件(Event),指在特定时间和环境下发生的,由若干角色参与的,表现出一些动作特征的一件事情<sup>[17]</sup>.

事件的六元组表示, $e = (o, a, t, v, p, l)$ ,定义为一个 6 元组结构.其中,事件六元组中的元素称为事件要素,分别表示对象、动作、时间、环境、断言、语言表现.

定义 2. 事件类(Event Class),指具有共同特征的事件的集合,用 EC 表示:

$$\text{用 EC 表示: } EC = (E, C_1, C_2, \dots, C_6) \quad (1)$$

其中, E 是事件的集合,称为事件类的外延; O, A, T, V, P, L 称为事件类的内涵,分别是 E 中的每个事件在对应要素上具有的共同特性的集合.

定义 3. 因果关系(causality),若事件  $e^i$  的发生导致了事件  $e^j$  的发生,其因果关系度大于给定的阈值时,则称两事件间具有因果关系,表示为  $RCE(e^i, e^j)$ ,因果关系可分为显式因果关系与隐式因果关系.

定义 4. 显式因果关系(Explicit causality),指文本中事件与事件间存在连接词,可将事件间的关系明确表述出来,简称显式关系.

定义 5. 隐式因果关系(Implicit causality),指文本中事件与事件间缺少显式连接词,需要根据上下文推测事件间的语义关系类型,简称隐式关系.

定义 6. 事件要素间因果关系,描述事件要素间的因果关系,是事件内部要素因果结构的一种反映,揭示事件内部要素间的关系,是判断事件间因果关系的重要部分.

#### 3.2 事件及事件要素识别

为识别事件间的因果关系,需从文本中获取事件源.在实验室的前期研究工作上,本文通过文本结合上下文特征,把触发词、与触发词相近的词以及这些词的词性、位置信息、依存关系等作为一个整体来考虑,并根据这些特征来识别事件.把事件识别的问题当作事件分类的问题,借鉴文本分类的思想,判断包含触发词的句子是否能划归于某一类事件.采用 SVM(支持向量机)机器学习算法实现事件识别并按主题分类,同时在识别过程中采用基于规则的方法将事件的六要素进行补齐,为后续工作提供基础.

#### 3.3 候选关联事件抽取

针对文本中抽取出来的主题事件需要判定事件对的相关性,因为只有具有相关性的候选事件对才有可能有因果关联,同时能有有效的判定候选事件对,从而避免事件对的盲目结合.在获取候选事件中,假设文本空间  $D = \{d_1, d_2, \dots, d_n\}$  上有事件类主题集合  $T = \{t_1, t_2, \dots, t_n\}$  和事件集合  $EC = \{e_1, e_2, \dots, e_i, e_j, \dots, e_n\}$ ,则在事件类主题  $t_k$  中,设  $e_i, e_j$  为两个任意不相交的事件,即  $e_i \cap e_j = 0$ ,其中  $e_i, e_j \in EC, EC \subset T$ .它们之间的关联程度可用支持度来衡量,支持度有助于确定事件在给定文本句子集合  $n$  中的频繁程度,可用来衡量  $e_i, e_j$  间关系的强度.其具体表示如下.

$$s(e^i, e^j) = \frac{\partial(e^i \cup e^j)}{n(\partial)} \quad (2)$$

其中  $s(e^i, e^j)$  表示事件  $e_i, e_j$  的相关度,  $n(\partial)$  表示文本语

料集中包含的句子集合.

当  $P(e_i, e_j | t_k) = s(e^i, e^j) > 0$  时, ( $\theta$  为主题相关度阈值, 在本文  $\theta$  值取 0.02) 认为  $e_i$  与  $e_j$  具有相关性.

#### 4 因果关系共性特征

根据哲学、逻辑学和统计学等领域对因果关系理论可知, 因果关系实质是两个事件之间的原因和结果有向的关系, 在文本事件中可理解为两个事件要素间的逻辑或定性作用机理的描述. 不管显式因果事件还是隐式因果事件, 它们遵循以下几个共性特征:

① Hume 的哲学规律理论<sup>[18]</sup>, 时间上原因事件发生在结果事件前, 因果事件是指两个事件间的固有的联系, 在时间上相近, 在空间上相邻.

② 结果事件依赖于原因事件, 在哲学和逻辑学中因果关系操纵理论 (manipulation theory of causality)<sup>[19]</sup> 将事件的因果关系看成操纵与控制目的的关系, 即如果  $e_i$  是  $e_j$  的原因, 那么  $e_i$  在一定的合适的方式控制来操纵或改变  $e_j$ .

③ 因果关系可通过概率或者分布函数的角度体现出来的. [Suppes] 等<sup>[20]</sup> 认为: 如果事件  $e_i$  的发生增加了事件  $e_j$  发生的概率, 那么事件  $e_i$  就是事件  $e_j$  的原因. 尽管 Suppes 的理论没有考虑事件发生的先后导致出现观测性等价问题, 使统计上无法区分原因和结果, 但从另一方面表现出了事件成对出现的概率大于事件单独出现的概率.

#### 5 事件间因果关联度计算方法

针对文本中所获取的候选关联事件对, 结合前文所描述的因果关系特点可知, 判定事件间因果关联度的实质是判定事件六要素之间的关联度. 事件的动作要素作为事件的核心要素, 通常也称为触发词 (动词或动名词为主), 它能基本涵盖事件表达语义, 需要作为判定事件间因果关系的重要参数之一. 同时在判定事件间的因果关系时还要考虑事件间其它要素之间的关系. 为此, 本文将结合事件间因果关系及事件要素特点分别计算事件间触发词的关联度、事件触发词与另一事件其它要素的关联度, 及事件间除触发词外其它要素之间的关联度. 同时考虑到事件要素的语义性, 将事件要素进行同义词泛化至第三层一并纳入计算.

##### 5.1 事件触发词间的语义关联度

根据因果关系的特征之一, Suppes 提出“因果关系可能性理论”中强调如果事件  $e_i$  是  $e_j$  可能的原因事件, 那么  $e_j$  同  $e_i$  共同发生的概率高于事件  $e_j$  本身单独出现概率, 也就是说  $P(e_j | e_i) > p(e_j)$ , 等价于  $p(e_i, e_j) / p(e_i)p(e_j) > 1$ , 这类似于计算  $e_i$  与  $e_j$  的 PMI<sup>[18]</sup> 值. 考虑到事件的动作要素是表达事件状态的重要成分, 也是判定事件因果关系的主要对象. 因此将上述的“因果关系可能性理论”通过计算  $e_i, e_j$  的动作要素  $a^i, a^j$  的 PMI 值表示.

由于互信息方法存在没有考虑特征在文本中发生的频度的缺点, 互信息是二元组  $P(x_i, x_j)$  的概率和两单独词的概率乘积  $P(x_i) * P(x_j)$  的似然对数比, 那将出现低频率词组成的二元组的互信息会大于高频率词组成的二元组现象, 这显然不能接受的, 为此, 本文采用改进后的 PMI' 来计算事件间触

发词语义关联度:

$$Ra^i a^j(e^i, e^j) = PMI'(a^i, a^j) = \log \left[ \frac{P(a^i, a^j)}{P(a^i) \times P(a^j)} \right] / \log \left[ \frac{2}{P(a^i) + P(a^j)} \right] \quad (3)$$

其中  $Ra^i a^j(e^i, e^j)$  表示事件对  $e_i, e_j$  动作要素  $a^i, a^j$  的关联度,  $P(a^i, a^j)$  表示  $a^i, a^j$  共同出现的概率,  $P(a^i)$  表示  $a^i$  出现的概率,  $P(a^j)$  表示  $a^j$  出现的概率, 如果  $P(a^i, a^j) < P(a^i) * P(a^j)$ , 则令  $P(a^i, a^j) = P(a^i) * P(a^j)$ , 这样有  $0 \leq PMI'(a^i, a^j) \leq 1$ .

改进的  $PMI'(a^i, a^j)$  可有效改变由于低频率词组成的二元组的互信息要大于高频率词组成的二元组的问题, 一定程度上弱化了由于数据稀疏对最终相关关系量化值的影响.

##### 5.2 事件触发词与事件其它要素的语义关联度

事件间的关联度除了同事件间的动作要素息息相关外, 还同事件的其它要素相关. 如“车祸发生后, 医务人员迅速赶往现场”, 这句话中可发现“车祸”事件的对象要素 (主体: 伤亡者), 地点要素, 时间要素, 与“赶往”事件的对象要素 (客体: 伤亡者), 地点要素 (同“车祸”事件的地点要素), 及时间要素等都是关系紧密的. 因此事件的因果关系可看成事件间各要素间的操纵与控制目的的关系, 假设  $e_i$  是  $e_j$  的原因, 那么  $e_i$  的要素在一定的合适的方式下控制操纵或改变  $e_j$  的要素.

因此, 可通过等式 (4) 来衡量事件  $e_i(e_j)$  的触发词 (动作要素) 与  $e_j(e_i)$  动作要素外的其它要素的语义关联度, 等式 (5) 来描述  $e_i(e_j)$  间除动作要素外的其它要素的相互关联度的集合.

$$Ra\epsilon(e^i, e^j) = \frac{1}{|\epsilon^i|} \sum_{\epsilon^i} PMI'(a^i, \epsilon^j) + \frac{1}{|\epsilon^j|} \sum_{\epsilon^j} PMI'(a^j, \epsilon^i) \quad (4)$$

$\epsilon = (o, t, v, p, l)$

其中  $\epsilon^i(\epsilon^j)$  表示事件  $e_i(e_j)$  中除动作要素以外其它要素集合  $(o, t, v, p, l)$ , 等式 (4) 表示事件  $e_i(e_j)$  的动作要素与事件  $e_j(e_i)$  的其它要素的语义关联度集合.

$$R\epsilon^i \epsilon^j(e^i, e^j) = \frac{1}{|\epsilon^i| |\epsilon^j|} \sum_{\epsilon^i} \sum_{\epsilon^j} PMI'(\epsilon^i, \epsilon^j) \quad (5)$$

$\epsilon = (o, t, v, p, l)$

其中  $\epsilon^i(\epsilon^j)$  表示事件  $e_i(e_j)$  中除动作要素以外其它要素集合  $(o, t, v, p, l)$ , 等式 (5) 表示事件  $e_i, e_j$  除动作要素外的其它要素的相互关联度的集合.

##### 5.3 事件间依赖度

接下来根据因果关系操控理论的特点, 即结果事件依赖于原因事件, 也就是结果发生依赖于原因. 这种依赖性能从因果性可能性理论的计算中体现出来, 但缺乏相互依赖程度的度量. 而这种度量可以映射到语句间的距离, 如果两个语句相关程度高, 则之间的距离接近, 反之疏远. 从事件的角度来讲, 句子的距离同句子间的关系度是成反比的, 也就是说相邻的句子间的依赖程度高于相隔较远的句子间的依赖程度. 为此可在对 Leacock and Chodorow<sup>[21]</sup> 提出的通过两节点之间的最短路径长度来衡量语义相似度方法略为改进, 用来判定事件对之间的依赖程度. 其计算  $e^i, e^j$  距离的公式如下:

$$Dist(e^i, e^j) = -\log \frac{sen(e^i) + sen(e^j) + 1}{2 \times ws} \quad (6)$$

其中  $sen(e)$  表示出现事件中动作  $v$  的句子数,  $ws$  窗口大



小定义为3,也就是说出现 $e^i, e^j$ 的句子的距离最多是3个句子,否则似为关系不紧密.当 $e^i, e^j$ 出现在同一句子时,上述的分子返回1.

5.4 事件间因果关系的度

最后,综上事件因果关系的特点对应的计算方法,可以得出 $e_i e_j$ 事件间因果关系的计算公式(RCE)为:

$$RCE(e^i, e^j) = Ra^i d^i(e^i, e^j) \times Ra^e(e^i, e^j) \times R^e e^j(e^i, e^j)$$

(7)

最后设定 $RCE(e^i, e^j)$ 的阈值作为判定事件 $e_i e_j$ 是否存在因果关系的二元分类器.

6 实验结果与分析

6.1 实验背景

为验证上述因果事件抽取方法的可行性,需要相应的语料库来进行训练与测试.在实验室的前期研究工作<sup>[22]</sup>中,将从网上收集了203篇关于地震、火灾、交通事故、食物中毒以及恐怖袭击等5类突发事件的新闻报道作为生语料(CEC语料库),根据前文所提的工作对语料经过分句、词性标注、抽取事件要素、去除否定事件、标记事件间的关系等预处理.在标注事件间的关系时,首先将标注人员分为两组,分别独立的对语料进行标注.允许标注明确的各类非分类关系,其中因果关系用RCE(含显式与隐式)表示,存在争议的因果关系事件用Rd表示,其它关系用R表示.标注人员将根据事件对的触发词自行识别并标注RCE、Rd、R关系,并自主在不限制固定窗口大小句子条件下对事件触发词标注关系.同时对于含有存在争议的因果关系事件,双方参与一起表决.

表1 CEC语料数据统计表/个  
Table 1 CEC corpus data

实验语料	文档	事件数	句子数	因果关系事件数		时间 短语数
				显式	隐式	
交通事故	54	837	265	97	123	203
地震	45	704	292	88	159	217
火灾	31	531	260	66	107	199
食物中毒	43	191	288	43	79	214
恐怖袭击	30	490	249	71	93	218
总计	203	2753	1354	365	561	1051

事实上,在本实验人工标注事件间的因果关系时,两组标注者对一些争议性的因果关系各抒己见,在5类文档中各随机抽取10篇文档让标注者进行重复标注.在相同文档中,在因果关系的识别中,两组标注者对RCE、Rd、R关系数标注的个数并不一致,分别是204、15、1210(第一组),191、16、1216(第二组),经过讨论一致同意196个为因果关系,其它关系数1208个.

最后大家在给定的200篇语料中标注出每个文档集中一致认可的事件项数目以及具有关系的事件对数目等信息统计如表1所示.为了确定 $RCE(e^i, e^j)$ 的阈值,需要根据训练语料不断调整校准它,使之在 $[0, 1]$ .为此,在训练语料库中计算那批重复标注的语料中具有关联事件对的 $RCE(e^i, e^j)$ 值并进行排序,并根据已标注的关系情况进行对比,查找一个最接近两组标注情况的 $RCE(e^i, e^j)$ 值作为认定事件因果与非

因果关系的阈值(0.87).

6.2 实验结果与评价

为了验证本文提出方法的有效性,选取近期国内外其它学者对抽取因果事件研究方法与本文方法进行实验对比进行评价.

首先Riaz and Girju<sup>[14]</sup>依赖于因果关系的操作原理和精度与召回性能曲线的比较并提出EDC方法来计算事件间是否存在因果关系.

其次是Quang Xuan Do<sup>[15]</sup>所提出的CEA方法,该方法是通过计算事件的谓词及论元的关系与距离来判定事件间是否存在关系,并联合篇章中的细粒度因果关系连接词来判定事件间因果关系.

接下来,再来评估B. Steven<sup>[10]</sup>提出的ML方法及本文所提出的RCE方法,通过训练集确定RCE的阈值后,用现有的CEC语库中的测试文档集进行验证.

实验中,本文从CEC语料库中的五个事件类中各随机抽取20篇,共100篇文本语料来对本文方法进行验证,类同前面学者的测评方法一样,通过计算本方法P、R、F三个指标的值进行比较.实验结果如表2所示.

表2 本文实验结果与其他方法研究结果比较  
Table 2 Results of this article compares with other methods

抽取方法	实验语料数据	实验结果		
		P	R	F
EDC <sup>[14]</sup>	760K documents in the English Gigaword corpus	0.409	0.235	0.297
CEA <sup>[15]</sup>	CNN of 2010, 25 articles as evaluation data	0.585	0.305	0.401
ML <sup>[10]</sup>	TimeBank corpus	0.469	0.594	0.524
RCE(本文方法)	从CEC语料库中五个事件类中各随机抽取20篇,共100篇文本语料	0.622	0.580	0.599

由表2可见,从不同方法下不同语料抽取的因果关系事件效果对比中可看出,本文提出的方法在召回率、准确率和调和平均值上略高于其它类别.

本文认为实验效果较好的主要原因是将文本划分为事件表示,一者能使事件描述更为明确,二者事件的六要素本身具有一定的语义成份,第三是事件的表示方法有助于发现关联事件的候选事件,为抽取因果关系事件打下良好的基础.事实上,文本用事件表示后本来就没有连接词的存在了,就事件而言,关系都是隐式的了.当然本文的效果也许同本实验的语料局限于突发事件类也有一定的关系.

7 结束语

具有显式因果连接词文本能直观的判定因果关系,但对于没有因果连接词的文本事件却难以判定,当前此类研究也较少.本文提出的基于事件语义的因果关系抽取方法既能针对文本中隐性的因果关系的抽取,对显式因果关系抽取同样有效.该方法将文本按事件划分,鉴定事件及事件的角色,通过计算事件的语义角色(六要素)间关联度来判定事件间的关联度;接着根据事件间因果关系存在时间优先性、因果概率评价性、因果关系依赖性等特性列举综合计算等式用来计算

候选事件间的因果关联度,并设定因果关系的阈值作为判定事件间因果关系的二元分类器.实验结果表明,本文提出的方法能够较好地抽取事件间的因果关系,无论语句间是否存在因果关系连接词,且不受语料特征的局限,同时抽取过程中该方法不限于句子间的因果关系,对跨句的因果关系抽取仍适用.但是实验未对因果关系事件对的原因事件及结果事件等具体关系进行抽取,这是后期需要将继续研究和解决的问题.

## References:

- [1] Girju R, Moldovan D. Text mining for causal relations[C]. In: Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference, Florida, USA: AAAI Press, 2002: 360-364.
- [2] Garcia D, COATIS. An NLP system to locate expressions of actions connected by causality links[C]. Proc of the 10th European Workshop on Knowledge Acquisition, Modeling and Management. Catalonia, Spain, 1997: 347-352.
- [3] Khoo C, Kornfilt J, Oddy R, et al. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing[J]. Literary and Linguistic Computing, 1998, 13(4): 177-186.
- [4] Blanco E, Castell N, Moldovan D. Causal relation extraction[C]. Proc of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco, 2008: 310-313.
- [5] Gan Hong-hua, Pan Yun-he. A new analysis of the structure of event causation[J]. Pattern Recognition and Artificial Intelligence, 2003, 16(1): 56-62.
- [6] Girju R. Automatic detection of causal relations for question answering[C]. In: Proceedings of the 41st ACL Workshop on Multilingual Summarization and Question Answering, Sapporo, Japan: ACL, 2003: 76-83.
- [7] Marcu D, Echihiabi A. An unsupervised approach to recognizing discourse relations[C]. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, USA: ACL, 2002: 368-375.
- [8] Inui T, Inui K, Matsumoto Y. What kinds and amounts of causal knowledge can be acquired from text by using connective markers as clues? Lecture notes in computer science[C]. Berlin Heidelberg: Springer-Verlag, 2003, 2843: 180-193.
- [9] Zhong Jun, Yu Long, Tian Sheng-wei, et al. Causal relation extraction of uyghur emergency events based on cascaded model[J]. ACTA Automatica Sinica, 2014, 40(4): 771-779.
- [10] Steven B, Martin James H. Learning semantic links from a corpus of parallel temporal and causal relations[C]. In Proceedings of ACL-08: HLT, Association for Computational Linguistics, 2008: 177-180.
- [11] Sun Y, Liu N, Xie K, et al. Causal relation of queries from temporal logs[C]. In Proceedings of WWW. Banff, Canada: ACM Press, 2007.
- [12] Zheng Jiao-lin, Tang Chang-jie, Qiao Shao-jie, et al. Mining causality in sub-complex dynamic system based on perturbation[J]. Chinese Journal of Computer, 2014, 37(12): 2548-2562.
- [13] Beamer B, Girju R. Using a bigram event model to predict causal potential[C]. In Proceedings of in Computational Linguistics and Intelligent Text Processing (CICLING-09), 2009.
- [14] Mehwish Riaz, Roxana Girju. Another look at causality: discovering scenario-specific contingency relationships with no supervision[C]. Proceedings of International Conference on Semantic Computing (ICSC'10), 2010: 361-368.
- [15] Quang Xuan Do, Yee Seng Chan, Dan Roth. minimally supervised event causality identification[C]. In Proceedings of Empirical Methods on Natural Language Processing (EMNLP'11), 2011.
- [16] Fu Jian-feng, Liu Zong-tian, Liu Wei, et al. Event causal relation extraction based on cascaded conditional random fields[J]. Pattern Recognition and Artificial Intelligence, 2011, 24(4): 567-573.
- [17] Liu Zong-tian, Huang Mei-li, Zhou Wen, et al. Research on event-oriented ontology model[J]. Computer Science, 2009, 36(11): 189-192, 199.
- [18] Hume D. A treatise of human nature[M]. Oxford: Oxford University Press, 1978.
- [19] Woodward J. Making things happen: a theory of causal explanation[M]. Oxford: Oxford University Press, 2003.
- [20] Suppes Patrick, A probabilistic theory of causality[M]. Amsterdam: North-Holland Pub, 1970.
- [21] Leacock C, Chodorow M. Combining local context and wordNet similarity for word sense identification[M]. WordNet: An Electronic Lexical Database, 1998: 265-283.
- [22] Fu Jian-feng, Liu Wei, Liu Zong-tian. A study of Chinese event taggability[C]. Proc of the 2nd International Conference on Communication Software and Networks. Singapore, Singapore, 2010: 400-404.

## 附中文参考文献:

- [5] 干红华, 潘云鹤. 一种基于事件的因果关系的结构分析方法[J]. 模式识别与人工智能, 2003, 16(1): 56-62.
- [9] 钟 军, 禹 龙, 田生伟, 等. 基于双层模型的维吾尔语突发事件因果关系抽取[J]. 自动化学报, 2014, 40(4): 771-779.
- [12] 郑皎凌, 唐常杰, 乔少杰, 等. 基于扰动的亚复杂动力系统因果关系挖掘[J]. 计算机学报, 2014, 37(12): 2548-2562.
- [16] 付剑锋, 刘宗田, 刘 炜, 等. 基于层叠条件随机场的事件因果关系抽取[J]. 模式识别与人工智能, 2011, 24(4): 567-573.
- [17] 刘宗田, 黄美丽, 周 文, 等. 面向事件的本体研究[J]. 计算机科学, 2009, 36(11): 189-192, 199.