相关事件挖掘与角色联系发现的研究

彭会良[1,2] 曹存根[2]

(首都师范大学计算机科学联合研究院 北京 100037)[1] (中国科学院计算技术研究所 北京 100080)[2]

摘 要 许多研究人员认为人们是以事件为单位来体验和认识世界的。以动词为核心的事件,把实体概念有机地联系、组织起来,在丰富实体概念间静态联系的同时,也构成了用以表示动态过程的基本单元。但是,事件知识却不容易从文本中直接获取。提出了一个以一个事件作为核心挖掘与之相关联的事件的方法。该方法在充分利用句法分析的基础之上,从二元词语扩展到语义更丰富的多个词语,挖掘到了相关的事件短语。在此基础上,机器标注了事件短语的人物角色,最终发现了相关事件与核心事件间的角色联系。实验结果显示,提出的方法从受限的文本语料里得到了大量的相关事件和角色联系,并取得了较高的准确率。

关键词 事件,相关事件,角色联系,知识获取

Research on Mining of Associated Events and Discovery of Roles Relationship

PENG Hui-liang[1,2] CAO Cun-gen[2]

(Joint Faculty of Computer Scientific Research, Capital Normal University, Beijing 100037, China)[1]

(Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)[2]

Abstract Many researchers hold this option that event accords with human normal cognitive rules and is the basic unit of human cognition. It is event which uses a verb as event core that constitutes the basic unit of dynamic process description while putting the entity concepts into organic links and organization to enrich static contact between them. But knowledge of event can not easily and directly acquire from text. A method based on one event as core event to mine events which are associated with the core event was proposed. It acquired associated event's chunks by extend binary words to multiple more semantic words on the basis of the full use of parsing, and then tagged role of the event's chunk and finally discovered roles relationship between of associated event and core event. The experimental results show that the proposed method has acquired considerable associated events and roles relationships from restricted corpus and has achieved a higher precision.

Keywords Event, Associated event, Roles relationship, Knowledge acquisition

1 引言

机器可读的知识是人工智能的基础。而如何利用计算机从文本中自动获取知识,是知识工程领域面临的难题之一[1]。文本挖掘,或文本数据库中的知识发现(Knowledge Discovery from textual Database)[2]是从大量文本数据中提取以前未知的、有用的、可理解的模式或知识的过程。传统的基于向量空间模型(VSM)的挖掘方法,由于没有利用文档的语义信息,难以达到更高的准确度。也有很多研究者对VSM进行了各种改进,引入了更多的语义信息,但目前仍缺乏系统的方法来对语义信息进行获取和表示。

很多研究人员认为,人类是以"事件"为单位来体验和认识世界的,事件符合人类的正常认知规律。然而,对事件的研究国内外都还刚刚起步。Nelson从认知科学上提出了事件表示模型[3],将事件表示作为分类知识的基础;他认为事件是对象和关系的一个整体,包括为特定目标的行为人对物体的作用,以及人们相互之间的作用。上海大学的周文提出事件多元组模型[4],在这种模型里,事件包括动词和动词连接的高频名词或者命名实体,一个事件可以表示为一个多元组:$\{n_1, \cdots, n_i, v\}$,该方法将文本断句、分词、句法分析后,识别出命名实体和高频名词,把核心动词作为该事件的核心行为词,得到一个事件多元组,从而保留了较多的语义信息。清华大学吴平博、湖南大学梁晗等采用基于框架法的事件信息提取技术[5,6],用预先定义的框架从语料中提炼事件,对灾难性事件的不同侧面进行表示;这种方法在事件抽取时,绕过了句法分析,降低了实现的难度。中国科学院计算技术研究所姜吉发等通过事件模式、事件触发模式和事件抽取模式定义事件框架,通过关键词来定位事件的候选描述语句[7]。

在事件的相互关联中,Talmy对包含因果关系的事件提出因果事件框架(Event Frame)来进行分析[8]。框架细分为5

到稿日期:2010-01-08 返修日期:2010-03-21 本文受国家自然科学基金(60573063,60773059)和中国高技术研究发展计划(863)(2007AA01Z325)资助。

彭会良 男,硕士,主要研究方向为文本挖掘、知识获取,E-mail:pengwei1_2000@163.com;**曹存根** 研究员,博士生导师,主要研究方向为人工智能、大规模知识处理。

个展示形成整个因果过程的不同阶段的次事件；任一后续次事件均蕴含其前次事件，中间事件可以省略和重合。5 个次事件是吸引对事件注意力的视窗（Attentional Windows），可开启，但不必全部开启就可实现完整表义。有的学者使用元组（有限元素的有序集）表示事件之间的关联[9]，例如：批评过程=（生气，指出错误，改正）；或者利用领域过程树模板表示事件知识[10]：（领域过程树（领域）（主事件）（必需事件）（可能事件））。话题跟踪与检测（Topic Detection and Trac-king，TDT）课题主要研究不同新闻文本如何有效地组织和搜索与结构化[11]。学者给出了事件检测与事件关系发现的概念，事件检测是在一个话题内对事件进行建模，事件关系发现是对话题内的事件依赖关系进行建模[12]。句法分析技术根据给定的语法体系，自动推导出句子的语法结构，分析句子所包含的句法单位和这些句法单位之间的关系[13]。通过句法分析可以将句子由一个线性序列转化为一棵结构化的依存分析树，通过依存弧反映句子中词汇之间的依存关系。

本文第 2 节提出了多元组事件的定义，明确了所研究的问题；第 3 节介绍如何抓取事件语料；第 4 节、第 5 节分别对相关事件的挖掘、角色联系发现进行了讨论；第 6 节分析了相关的实验结果；第 7 节讨论了本文提出的方法的适用情况。

2 问题定义

本文吸纳了多元组模型的定义，将事件定义如下：事件 $EV=(not,core,N,A,R)$ 是一个五元组，其中：

1）*core* 是事件的一个核心词，一般是动词或形容词，不可缺省。

2）*not* 是事件中的一个否定性副词，表示事件是否发生，可以缺省。

3）$N=\{n_1,\cdots,n_s\}$ 为非空集合。其中的元素为事件 *EV* 涉及的名词或命名实体。这些名词或命名实体充当整个事件 *EV* 的事元。所谓事元，就是构成事件的直接论元[14]。彼此互不相同的事件含有的事元的种类和数量也不相同。在本文里，我们将这些事元不加区分地都称为事件角色，将其中由人来担任的事件角色称为角色人，这些名词或命名实体称为角色词。

4）$A=\{a_1,\cdots,a_z\}$，可以是空集；非空时进一步说明事件的程度、结果等内容，多为形容词。

5）$R=\{r_1,\cdots,r_k\}$，可以是空集；非空时标识出这个事件中的角色与其他事件的角色之间的联系，即角色联系。

下面举例说明这种联系，先给出几个句子：a. 苹果熟了；b. 他们摘下苹果；c. 他们批评了这个同学；d. 他哭了。上面的每个句子表述了一个事件，在句子 a 和 b 中，词语“苹果”可以把 a 和 b 联系到一起：苹果熟了，人们会把它摘下来。类似于这样的联系是我们希望加入到 *R* 的。

当两个事件包含着一定的语义逻辑，反映某些知识联系时，称这两个事件互为相关事件。a 和 b 表述的事件就是一对相关事件。

然而，当用不同事件中相同或相近的角色词来简单、直接地建立不同事件之间的一个角色联系时，这两个事件却不一定是相关事件。比如：句子 b 和 c 表述的事件中有相同的词“他们”，但 b 和 c 在语义逻辑上并没有明显关联。这样的联系就不能加入 *R*，是我们所不希望标识的。c，d 中虽然没有相同的角色词，而一个人挨了批评而哭泣是很符合常理的，“他”也可能就指代了“这个同学”，这个联系要用 *R* 的元素来加以标识。从这几个例子中看出，由于代词的大量使用，当通过事件的角色人来建立角色联系时，角色联系发现的难度就大大增加了。

角色联系是由 *R* 中的元素来表示的。设 e_i 是一个多元组事件，当 e_i 的 *R* 是非空集时，*R* 中的一个元素 r_i 是一个角色对：$\langle n_g,e_j::n_f\rangle$。其中，$e_j$ 表示不同于本事件 e_i 的另一个事件；n_g 表示本事件的一个角色 n_g；$e_j::n_f$ 表示事件 e_j 的一个角色 n_f。r_i 表示在 e_i 和 e_j 互为相关事件的情况下，本事件的一个角色 n_g 可以作为另外一个事件 e_j 的角色 n_f。当然 e_i 可以通过 *R* 中的元素标识出与其它事件的角色联系。

事件在文本里由事件短语来表述。比如句子 c 表述了一个人批评另一个人的一个事件，即一个“批评”事件，我们将 c 称为“批评”事件的事件短语。所谓事件短语指的是一个以谓语为核心、语义比较完整的句法单元。事件短语承载事件，事件是事件短语的内容；不同的事件短语把同一个事件以一种或几种依存分析树的形式表达出来，以便于适应语序的灵活性。事件短语不能简单地等同于简单句，它可以是一个简单句，也可以是复杂句里的一个语义完整的短语。事件短语最少由两个词语构成。比如：在句子“他打开门之后，就走出了教室”中，可以抽取出两个事件短语：1.“他打开门”；2.“[他]走出教室”。同时可以看到这两个事件短语都能构成一棵主＋谓＋补＋宾的句法树。在事件短语中我们更关心对表述事件至关重要的词语而往往忽略某些词语。比如在 c 句子里，“他们”、“批评”、“同学”3 个词就足以表达其主要意义，而“了”、“这个”被忽略了。从这个角度上，句子 c 作为一个事件短语包含 3 个词语。

假设 c 表述的“批评”事件是一个已知事件，用 e_c 标记。根据事件定义，词语“批评”作为事件的核心词，用 *kevent* 标识，忽略“批评”的具体人物，抽象出“批评者”、“被批评者”两个角色，分别用 *kx*，*ky* 标识，那么这个事件可以表示为：$e_c=(core:kevent,not:null,N:(kx,ky),A:(null),R:(null))$。其中，*null* 表示集合是空集或该元组是缺省的，$N:(kx,ky)$ 即事件的两个角色。句子 d 表述了一个与 e_c 相关的事件。把 d 中的“他”抽象为“某人”，根据事件的定义，d 表述的事件表示为：$e_d=(core:哭,not:null,N:(某人),A:(null),R:(\langle 某人,e_c::ky\rangle))$。其中，角色对 $\langle 某人,e_c::ky\rangle$ 标识出了这样一个联系：被批评的人哭泣。与已知的事件 e_c 相关的事件 e_d 就是本文最终想要获取的。

本文把由人作为事件实施主体的行为类事件作为研究的对象，希望挖掘出一个已知事件的相关事件，以及在这个已知事件中不同的角色人又分别参与了哪些相关的事件。这也就确定了挖掘过程的两个步骤：挖掘相关事件、发现角色联系。

下面以上文提到的“批评”事件为例简要说明挖掘的整个过程，同时给出本文后面使用的几个概念。我们又把已知事件“批评”称为核心事件，用 e_0 标记，那么 $e_0=(core:kevent,not:null,N:(kx,ky),A:(null),R:(null))$。设 $D=\{d_1,d_2\cdots,d_n\}$ 为包含 e_0 的上下文语料，每个 d_i 都是一段含 e_0 的文本。我们认为与 e_0 相关的事件大多出现在 e_0 前后的几十个词的范围之内，因而将 e_0 前后总计 50 个词作挖掘相关事件的范围，将 50 个词称为相关长度。把包含核心事件 e_0 且符

合相关长度限制的文本称为核心事件短文本，简称短文本。那么每个文本 d_i 的词语长度要符合相关长度的限制，每个 d_i 也就是一个核心事件短文本。在对事件语料 D 进行分词、依存句法分析等处理的基础上，按照本文提出的二元词语到多词的事件短语的扩展方法，在事件语料 D 中标出事件短语。我们把文本里连续的 8 个词作为标注事件短语的最大长度，将其称为事件长度。

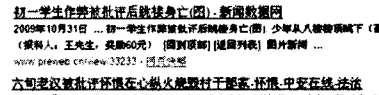
所谓的二元词语满足以下 3 个条件：1)二元词语包含不同的两个词，并且这两个词都曾在事件语料 D 中出现；2)两个词之间是有序的，即(被、批评)和(批评、被)是不同的二元词语；3)在事件语料 D 的依存句法分析结果中，二元词语间最少存在一次词汇依存关系。

标出事件短语后，由不同的事件短语得到 e_0 的相关事件的集合 $E=\{e_1,e_2,\cdots,e_m\}$，其中每个 e_i 表示一个与 e_0 相关的事件。最后，进一步去发现 e_i 与 e_0 之间的角色联系，确定核心事件中“批评者”、“被批评者”又分别参与了哪些相关的事件。

3 语料抓取和预处理

为了获取大量的包含 e_0 的事件语料，我们要先构造出表述 e_0 的事件短语，把这些事件短语作为搜索引擎的搜索项，然后通过抓取搜索项的返回页面来获得事件语料。下面，以“批评”事件为例说明语料抓取过程。

对“批评”事件，我们要求其事件短语仅仅表述了某一个人批评了另外一个人这一种含义，要努力做到避免产生歧义。把事件短语稍加改动，用“＊”代替部分词语，得到“被＊批评”、“把＊批评了”、“严肃地批评了”、“批评了我”等共 15 条短语，再与使用最广的普通词汇“的”、“了”、“人”等 30 个词搭配组合成 450 个搜索项。搜索引擎返回的片段如图 1 所示。


初一学生作弊被批评后跳楼身亡(图)-新闻数据网
2009年10月31日 ... 初一学生作弊被批评后跳楼身亡(图) 少年从八楼楼顶跳下(画圈处)。 ...
(据报人：王先生，奖励60元) [图到顶部] [返回列表] 图片新闻 ...
www.prweb.cn/new-33231 - 网页快照

六旬老汉被批评怀恨在心纵火烧毁村干部家_怀恨_中安在线_法治
2007年8月8日 ... 六旬老汉被批评怀恨在心纵火烧毁村干部家 ... 六旬老汉被批评怀恨在心纵火烧毁村干部家 2007年08月04日10时04分来源： 京华时报 ...
law.anhuinews.com/system/2007/08/08/001817254.shtml - 网页快照

图 1 搜索引擎返回项截图

从图 1 的两个返回项的网页正文里抓到了被“…”分割的不重复的两个网页片段：从第一个返回项抽到片段“初一学生作弊被批评后跳楼身亡(图). 少年从八楼楼顶跳下(画圈处)。”；从第二个返回项抽到片段“六旬老汉被批评怀恨在心纵火烧毁村干部家。2007 年 08 月 04 日 10 时 04 分来源：京华时报”。将网页片段保存为一个文本，共抓得不重复文本 56729 个。每个文本包含从十几个到几十个词不等，基本符合相关长度的限制。使用哈工大信息检索研究室提供的语言技术平台共享包进行分词和词性标注、依存句法分析，每个文本得到一个 d_i，得到了“批评”事件的语料 $D=\{d_1,d_2,\cdots,d_n\}$。

4 相关事件的挖掘

相关事件是通过事件短语来表述的，需要先标记出事件短语。标记过程是基于依存句法分析完成的。然而，目前的依存句法分析结果：1)整句分析的正确率很低，一个词的错误导致整个句法树错误，复杂句、长句正确率更低；2)简单句和长度较小的短语分析正确率较好，特别是比较规范的词语和常用的句法树结构。本文基于以上情况提出了从依存句法分析的结果中标注出事件短语的方法。由事件语料获取部分可知，本文使用的事件语料 D 是包含 e_0 的受限语料，各个短文本间存在一定的相似性。当某个与 e_0 相关的事件 e_i 反复出现、多次进行依存句法分析时，组成 e_i 的两个词之间所有的词汇依存关系中，越是频繁的词汇依存关系，就越与表述 e_i 的事件短语中真实的词汇依存关系趋向一致。由两个词的最频繁词汇依存向多个词的依存句法树扩展，纠正被错误分析的词汇依存关系，从而标注出以依存句法树的形式表现出来的事件短语，达到尽可能识别出表述事件 e_i 的事件短语的目的。

4.1 二元词语特征

二元词语是标注事件短语的基本单元，也是向多个词的事件短语扩展的起点。为了找到最合适的扩展起点，使标注出的事件短语与 e_0 相关性更强，需要对二元词语进行定量度量，通过观察语料，下面给出了刻画二元词语的几个特征。以下几个特征中，设词语 x,y 是在 D 中出现的不相同的两个词，而且 x 在 y 前，两个词在 D 中有至少一次的词汇依存关系，即词语 x,y 满足前文所提的条件，是一对二元词语。

特征一　二元词语外部点互信息

这个特征反映二元词语与 e_0 的点互信息(Point-wise mutual information)。点互信息越大，二元词语与 e_0 越相关。如式(1)所示：

$$MI'(xye_0,D)=\log_2\frac{P(xy|e_0)}{P(xy)} \tag{1}$$

式中，$P(xy|e_0)$表示在事件 e_0 出现的情况下，e_0 前后的相关长度内，x,y 以二元词语中的次序出现在同一个事件长度内的概率；我们把在事件语料 D 中，有 x,y 出现在同一个事件长度内，且与二元词语次序相同的短文本总数与 D 中短文本总数的比值作为 $P(xy|e_0)$的值；$P(xy)$表示在任意一个相关长度内，x,y 以二元词语中的次序出现在同一个事件长度内的概率；我们把普通语料库中 x,y 满足以下两个条件时，在同一个相关长度内同现的概率作为 $P(xy)$：1)x,y 在同一个事件长度内，且与二元词语同序；2)满足条件 1)的情况下，x,y 在同一个相关长度内多次出现，视为出现 1 次。最后通过式(2)做归一化处理，得到二元词语外部点互信息：

$$MI(xye_0,D)=\frac{MI'(xye_0,D)}{MI'_{\max外}} \tag{2}$$

式中，$MI'_{\max外}$是所有 $MI'(xye_0,D)$中的最大值。显然该特征中 $MI(xye_0,D)$不一定等于 $MI(yxe_0,D)$。

特征二　二元词语的 TFIDF

每个词语在事件语料 D 中的重要程度不同，从而引入了 TFIDF[15]。先给出一个词语 t 的 TFIDF 值的计算方法：

$$TI_1(t,D)=tf(t,D)\times\log_2(WEB/WEB_t) \tag{3}$$

式中，词语 t 是在 D 中出现的一个词，$tf(t,D)$为词语 t 在 D 中的出现频率，WEB 为所有 Web 页面的总数，该数字可取一个固定值；WEB_t 为所有 Web 页面中出现 t 的页面数，具体值是搜索引擎查询返回的含有该词的页面数目。直观而言，$TI_1(t,D)$值越大，这个词对 e_0 越重要，无非有两种可能：要么这个词构成了 e_0 事件本身，要么构成了 e_0 的相关事件。在式(3)的基础上，两个词的 TFIDF 值均采用 x,y 各自的 TFIDF 值的调和平均值(Harmonic Mean)计算，如式(4)所示：

$$TI_2(xy,D)=\frac{TI_1(x,D)TI_1(y,D)\times 2}{TI_1(x,D)+TI_1(y,D)} \quad (4)$$

如果两个词各自都有很高的 TFIDF 值，但却很少出现在同一个事件长度内，他们形成同一个事件的可能性就较小。因此我们用词语 x,y 在事件长度内的同现率对式(4)进行修正。设词 x 在 D 中出现的总次数为 $S(x)$，词 y 在 D 中出现的总的次数为 $S(y)$，在 D 中 x,y 以二元词语的次序同时出现在同一事件长度内的次数记为 $M(x,y)$，x,y 在事件长度内的同现率 M_{xy} 计算方法如式(5)所示，显然，M_{xy} 不一定等于 M_{yx}。

$$M_{xy}=\frac{M(x,y)}{S(x)+S(y)-M(x,y)} \quad (5)$$

用式(5)将式(4)修正为式(6)：

$$TI_2(xy,D)=\frac{TI_1(x,D)TI_1(y,D)\times 2}{TI_1(x,D)+TI_1(y,D)}\times M_{xy} \quad (6)$$

这样，同现率大的两个词的权重就得到了加强。最后用式(7)进行归一化处理，得到归一化的二元词语 TFIDF 值：

$$TFIDF(xy,D)=\frac{TI_2(xy,D)}{TI_{2_{\max}}} \quad (7)$$

式中，$TI_{2_{\max}}$ 是所有 $TI_2(xy,D)$ 中的最大值，这里采用归一化方法使得一个词的权重变化范围在 0－1 之间，不同于传统的把所有词的权重和作为 1 的归一化方法。这个特征中 $TFIDF(xy,D)$ 不一定等于 $TFIDF(yx,D)$。

特征三　二元词语内部点互信息

这个特征反映二元词语 x,y 彼此间的点互信息。二元词语内部点互信息的计算方法如式(8)所示：

$$MI'(xy,D)=\log_2\frac{(F(xy)+0.5)\times n}{(F(x)+0.5)(F(y)+0.5)} \quad (8)$$

式中，$F(x)$，$F(y)$ 分别表示 D 中包含 x 或 y 的短文本数，$F(xy)$ 表示在 D 中词 x,y 同现的短文本数，n 表示 D 的短文本总数，0.5 是数据平滑因子。然后对二元词语内部点互信息做归一化处理，计算方法如式(9)。这个特征中 $MI(xy,D)$ 等于 $MI(yx,D)$。

$$MI(xy,D)=\frac{MI'(xy,D)}{MI'_{\max内}} \quad (9)$$

式中，$MI'_{\max内}$ 是所有的 $MI'(xy,D)$ 中最大的点互信息值。

特征四　二元词语词汇依存强度

在事件语料 D 的词汇依存关系统计结果中，二元词语一般会有多种词汇依存关系，下面表 1 是“批评”事件语料的部分统计数据。

表 1　二元词语的词汇依存强度

二元词语(xy)	工作|认真	认真|工作	学|不好	犯|错
ATT 依存	0	1	1	7
ADV 依存	0	5	0	2
CMP 依存	1	0	17	14
SBV 依存	11	0	0	0
VOB 依存	0	0	0	227
依存总和(psum)	12	6	18	250
窗口同现(S)	27	18	25	827
最大依存(p)	SBV:11	ADV:5	CMP:17	VOB:227
最大依存比，式(10)	0.88	0.77	0.92	0.91
依存同现比，式(11)	0.44	0.32	0.71	0.30
依存强度，式(12)	0.59	0.45	0.80	0.45

如：二元词语(犯、错)有状中结构(用 ADV 标记，下同)、定中关系(ATT)、动补结构(CMP)、动宾关系(VOB)等 4 种词汇依存关系[16]。这里取其中最频繁的依存关系 VOB 作为(犯、错)唯一的词汇依存关系，并将被依存的词语“犯”称为“错”的依存前驱词。在“犯|错”中，符号“|”表示二元词语在句子里不一定紧密相邻，可能会插入其他成分。同样，表 1 中“工作|认真”最频繁依存是 SBV，“认真”是被依存词，即依存前驱词。通过 D 中词汇依存的统计，每个二元词语获得唯一的一个最频繁的词汇依存关系。下面，分别计算出二元词语最大依存比 $PARS_1(xy,D)$，如式(10)所示，依存同现比 $PARS_2(xy,D)$，如式(11)所示。

$$PARS_1(xy,D)=\frac{p}{psum+0.5} \quad (10)$$

$$PARS_2(xy,D)=\frac{psum}{S+0.5} \quad (11)$$

式中，S 是在 D 中 x,y 在事件长度窗口内以二元词语次序出现的同现频率(见表 1)，p 是 x,y 在事件长度窗口内最频繁的词汇依存关系的频率，$psum$ 是 x,y 在事件长度窗口内全部词汇依存关系之和，三者的数值满足 $0<p\leqslant psum\leqslant S$；0.5 是数据平滑因子。最后，二元词语词汇依存强度 $PARS(xy,D)$ 计算方法如式(12)所示。这个特征中 $PARS(xy,D)$ 不一定等于 $PARS(yx,D)$。

$$PARS(xy,D)=\frac{PARS_1(xy,D)PARS_2(xy,D)\times 2}{PARS_1(xy,D)+PARS_2(xy,D)} \quad (12)$$

小结：以上 4 个特征中，二元词语外部点互信息和二元词语的 TFIDF，倾向于度量二元词语与 e_0 联系的强弱程度，强调与 e_0 的相关性；二元词语内部点互信息和二元词语词汇依存强度，倾向于度量二元词语之间构成一个事件短语能力的强弱程度，强调二元词语的事件性。

4.2　相关事件挖掘算法

事件短语的标记是一个从两个词到多个词的扩展过程，应该从包含事件核心词的二元词语开始。由于某些动词用动宾短语表达一个核心语义，因而算法将有 CMP、动词作谓语的 SBV、形容词做谓语的 SBV、一价动词做谓语的 VOB 等 4 种词汇依存关系的二元词语确立为事件短语的扩展起点。在保留了哈工大句法分析已有的 CMP，VOB，SBV，POB[16] 的基础上，根据语义表达的完整性需要，从 ADV 中分离出了 TBA，BA，BEI，N-ADV 等 4 类新的词汇依存关系，从而得到了 8 种基本的词汇依存关系，新添加的词汇依存关系其意义如表 2 所列。

表 2　新添加的词汇依存关系

词汇依存关系标记	依存关系描述	依存关系的实例
N-ADV	否定词做状语	不|高兴
TBA	“向”类介词做状语	向|敬礼
BEI	“被”做状语	被|训斥
BA	“把”做状语	把|骂

受鲁川的汉语句模思想的启发[14]，语义完整的短语绝不是二元词语就能表达的。因而还需要总结从两个词的词汇依存到多个词的依存句法树的扩展规则。

下面举例说明这种规则，先给出几个句子：

1)“…他没有回答这个问题….”

2)“…由于没有回答出那个问题而挨了批评….”

3)“…他把这个问题回答出来了….”

4)“…会上，一共回答了四个问题….”

5)“…数学卷中，我把简单的问题也回答错了….”

6)“…没有把问题回答出来就跑了….”

虽然依存句法分析还不很准确，但全部 6 个句子的二元词语的词汇依存关系有助于我们发现“没有|回答|出|问题”这个事件短语。参照图 2，做出如下分析：

从 1)，2)，4)句都可以获取“回答|问题”的 VOB 依存；以 VOB 依存为扩展的起点，由 2)，3)的“回答|出/出来”得到 CMP 依存，在 2)中扩展为短语“回答|出|问题”，参见图 2 中粗箭头(1)；由 1)，2)，6)的“没有|回答”得到 N-ADV 依存，在 2)中扩展为短语“没有|回答|出|问题”，参见粗箭头(2)；我们将粗箭头(1)，(2)这样一条扩展路线称为合法路径。而 6)中的把字句则走了另外一条路径：从 2)，3)，6)都可以获取“回答|出/出来”的 CMP 依存；以 CMP 依存为起点，由 3)，6)的“把|回答”得到 BA 依存，在 6)中扩展为短语“把|回答|出来”，参见粗箭头(4)；由 3)，6)的“把|问题”得到 POB 依存，在 6)中扩展为短语“把|问题|回答|出来”，参见粗箭头(5)；由 1)，2)，6)的“没有|回答”得到 N-ADV 依存，在 6)中扩展为短语“没有|把|问题|回答|出来”，参见粗箭头(6)。

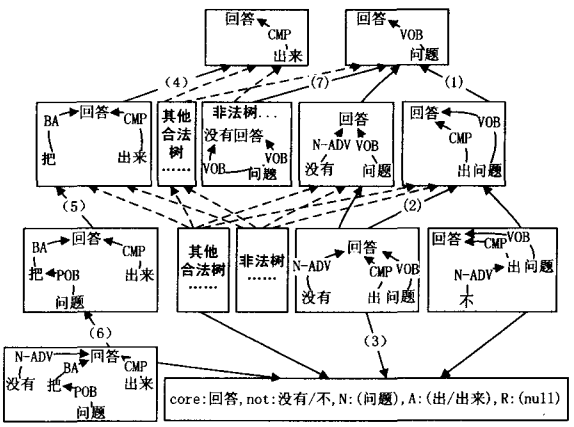


图 2　句法扩展树与句法扩展路径

扩展中还会产生非法树，所谓非法树就是在真实自然语言里找不到对应的依存句法树，它在语法逻辑下是不可能存在的。如：当我们从其他句子里获得“没有|问题”是 VOB 依存时，在 2)里扩展成“没有|回答|问题”短语，其扩展路径是粗箭头(7)，这次扩展所得结果就是一棵非法树。

可以估计出所有的不受句法逻辑限制的生成方法是庞大的，这其中有些不是树型依存，有些虽然是树形依存但不符合语法逻辑；句法扩展路径数目也相当庞大。我们通过人工标注事件短语内词语的词汇依存关系，利用算法自动生成并人工验证了 92 个使用率较高的合法扩展树以及他们之间的 1160 条合法扩展路径。

然而合法扩展树、合法扩展路径组成的句法树扩展规则依然会导致一个二元词语有多个合法的扩展。针对这个问题，提出了多个词的句法树依存强度的归纳计算方法。先看一个 3 个词的句法树依存强度计算的例子，设由式(12)已知：

$PARS$(回答|问题，D)=0.5，$PARS$(回答|出，D)=0.4；

则 3 个词的短语“回答|出|问题”的句法树依存强度为：

$PARS$(回答|出|问题，D)=0.5×0.4=0.2；其中“回答”是的二元词语“回答|出”中“出”的依存前驱词，也是二元词语“回答|出”挂到已有句法树“回答|问题”上时的挂载节点。“回答”、“出”、“问题”3 个词也是有序的，并且“出”在“问题”之前。将计算方法表示为：

$PARS(xzy,D)=PARS(xy,D)\times PARS(xz,D)$，其中 x 是 z 的依存前驱词，x 是挂载点，x,y,z 3 个词也是有序的，并且 z 在 y 之前。

依照例子，不难归纳出 $n+1$ 个词的句法树依存强度计算方法，如式(13)所示：

$$PARS(T,D)=\begin{cases}PARS(t_1\cdots t_n,D)\times PARS(t_it_{n+1},D)\\PARS(t_1\cdots t_n,D)\times PARS(t_{n+1}t_i,D)\end{cases}\tag{13}$$

式中，$1\leqslant i\leqslant n$，t_i 是本次扩展的挂载点，也是二元词语 t_it_{n+1} 或 $t_{n+1}t_i$ 中 t_{n+1} 的依存前驱词；T 表示 $n+1$ 个词的新序列。具体序列主要取决于 t_{n+1} 的插入位置，可能在 t_i 前面，也可能在 t_i 后面，同时还可能与 t_i 间隔一个或多个词(见图 2)。根据式(13)，在从 n 个词语向 $n+1$ 个词语扩展的过程中，挑选句法树依存强度最大的合法扩展作为本次扩展的唯一结果。

相关事件挖掘算法如下：

输入：集合 W，W'，Z，全部初始化为空集；事件语料 D；句法树扩展规则；二元词语外部、内部互信息预设阈值 $MI_{外}$、$MI_{内}$；

1)根据式(12)计算所有二元词语词汇依存强度：$Z=\{PARS(xy,D)\}$；根据式(7)计算二元词语的 TFIDF 值：$W=\{TFIDF(xy,D)\}$；根据式(2)和式(9)计算二元词语的外部和内部点互信息，将外部点互信息小于 $MI_{外}$ 或内部点互信息小于 $MI_{内}$ 的二元词语从 W 中删除。

2)在 W 中按照二元词语的 TFIDF 值从大到小，选择排名最靠前的二元词语加入 W'；在 W' 中辨别出可以作为扩展起点的二元词语，在 D 的短文本中，在事件长度窗口内找到对应的两个有序词，检查两个词的词汇依存关系是否与二元词语一致；当不一致时，改为与二元词语一致的依存关系。将获得的由两个词组成的事件短语，作为多个词的事件短语的扩展起点。把 D 中词汇依存关系变动更新到 Z。

3)再次按照 W 中的排名，增加 W' 中的二元词语；用 W' 现有的二元词语，添加新的扩展起点；从 D 中标出的两个词的事件短语开始，依照本文总结的句法树扩展规则找到所有合法扩展，根据 n 到 $n+1$ 个词的句法树依存强度计算式(13)挑选一个最大值作为唯一扩展结果；逐步扩展、完整标记出整个事件短语，同时将 D 内更改的词汇依存更新到 Z。

4)统计本次 W' 中新纳入的二元词语在 D 中对应的词对总数目，记为 q，在步骤 3)中被扩展成为事件短语一部分的词对数目，记为 p；当 $\frac{p}{q}$ 的值大于预设阈值 0.3，并且 W' 中二元词语总数与 W 中二元词语总数之比小于预设比例 0.6 时，返回到步骤 3)，否则中止。

5)根据 D 中扩展、标记出的事件短语，用同义词林合并相似短语。去除事件短语包含的介词，逐个生成相关事件；当一个事件对应的事件短语总数大于等于 3 时，把这个事件加入 E，得到 E。

输出：事件集合 E；每个事件对应的事件短语及对应的一种或多种依存句法树。

E 中记录了事件内容，事件短语的依存句法树则记录了事件的表述形式，图 2 底部就是一个多元组事件。在事件语料 D 中，事件短语里还夹杂着没有被抽取到事件内容中去的其他成分，如“犯了错误”、“犯了几次错误”、“犯了错”、“犯小错误”等。当把这些相关事件作为新的核心事件时，可以用这

些事件短语来获取事件语料。特别地，在5)中，当行为类事件作为核心事件 e_0 时，要对标出的事件短语中表示人的词语进行抽象，可以用“某人”来表示。比如：从“他|动|肝火”、“教授|动|肝火”抽象出相关事件“[某人]|动|肝火”。可以看出用“某人”处理比用“他”、“教授”等代词或社会角色词语更为合理。这样处理也便于进一步发现核心事件与相关事件的角色联系。

5 角色联系发现

角色联系发现的目标是发现核心事件中的不同的角色人及分别参与了哪些相关事件。代词指代消歧是文本处理的一大难题。e_0 的事件语料 D 中包含各类人物，增加了事件角色联系的发现难度。为了发现 e_0 与相关事件 e_i 的角色联系，我们通过观察语料，提出了以下假设：

1)D 中所出现的人物都是 e_0 中的角色人。即以“批评”为核心事件的短文本里，所有出现的人物只可能是“批评者”即 kx、“被批评者”即 ky 两种角色人之一。

2)短文本中，代词的指代具有稳定性，社会角色词有同指性。即人称代词、老师、爸爸等社会角色词指向同一个事件角色；如：在句子“他最终没有回答这个问题，李老师把他狠狠批评了一顿”里，识别出了两个事件短语：“他|没有|回答|问题”、字形式的 e_0 短语“把|他|批评”。字形式的 e_0 短语将被初始化为永久标记 kx，“把|他|批评”的“他”也被永久初始化为 ky。通过同指代词“他”，可以得到“他|没有|回答|问题”的角色“他”是一个“被批评者”，我们将把“他|没有|回答|问题”短语暂时标记为 ky。短语对应的事件的 ky 标记数就可以累加上1。

3)短文本中，叙述语句的主语有稳定性，文本总是显示或隐式地以某人的角度来推进；事件的角色转换有规律。比如下面是一个连动句“我没有回答出那个问题，被老师批评了”，句子里表述了两个事件，叙述主语相同。再比如下面的兼语句“老师批评我不认真”，表述了两个事件，事件“不认真”的我做了“被批评者”；事件角色的转换暗含在兼语句中。本文总结的角色转换规则，涵盖了连动句、兼语句、把字句、被字句、“向”类介词+角色+动词句、分句之间主语顺延、分句之间主语改变等规律。通过这些规律可以根据一个事件短语的永久角色标记对其紧邻的事件短语加注暂时角色标记。

在前面所举的连动句里被字形式的 e_0 短语将被初始化为永久标记 ky，“被|老师|批评”的“老师”也被永久初始化为 kx。由连动句的角色转换规则推出前面“没有回答出那个问题”的角色也是“被批评者”。因而将把“没有|回答|出|问题”短语暂时标记为 ky。短语对应的事件的 ky 标记数就可以累加上1。

对有两个角色人的行为类事件，如本文的“批评”事件，其相关事件 e_i 的角色人有两种可能，要么是 kx，要么是 ky。将两种可能的角色倾向分别记为 δ_{kx}，δ_{ky}，下面给出其计算方法：

$$\delta_{kx}=\frac{kx\text{标记数}+0.5}{ky\text{标记数}+0.5}\times\frac{kx\text{标记数}}{e_i\text{事件短语数}+0.5} \tag{14}$$

$$\delta_{ky}=\frac{ky\text{标记数}+0.5}{kx\text{标记数}+0.5}\times\frac{ky\text{标记数}}{e_i\text{事件短语数}+0.5} \tag{15}$$

式中，kx，ky 标记数分别为 e_i 的事件短语的两种暂时标记数，它们是通过累加假设2)和3)的分析中的暂时标记得来的，如表3所列；0.5是数据平滑因子。

表3 角色联系发现的部分数据

挖掘的相关事件（事件短语形式出）	事件短语数>3	暂时的kx标记	暂时的ky标记	人工标记的角色
[某人]|吸取|教训	45	3	10	被批评者
[某人]|违反|纪律	108	11	44	被批评者
[某人]|动|肝火	11	4	2	批评者
[某人]|没有|完成|作业	39	2	15	被批评者
[某人]|问明|情况	11	6	0	批评者
[某人]|下不来|台	11	2	3	被批评者

角色联系发现算法：

输入：角色联系事件集合 E_w，初始化为空集；相关事件集合 E；事件语料 D；e_0 的事件角色 kx，ky；根据假设3)总结的角色转换规则；角色倾向阈值 δ_0；

1)参照假设2)和3)，初始化“把”字形式的 e_0 短语、“被”字形式的 e_0 短语，分别加注永久的 kx 或 ky 标记。

2)对在 E 中而不在 E_w 中的每个 e_i，对其事件短语逐个分析、加注暂时角色标记：依照假设2)利用代词加注暂时标记；或者运用角色转换规则根据事件短语紧邻的其他事件短语的永久标记，加注暂时角色标记；达不到以上两个条件的不做标记。

3)运用式(14)、式(15)统计计算 e_i 的 δ_{kx}，δ_{ky}。当 δ_{kx} 或者 δ_{ky} 大于预设值 δ_0 时，用事件多元组中 R 的角色对记录下角色进行联系，并将 e_i 加入到事件集合 E_w；清除所有暂时标记；

4)将所有 E_w 中的事件对应的事件短语，都加注上永久标记 kx 或者 ky；

5)重复2)，直到 E_w 不再增大。

输出：角色联系事件集合 E_w。

6 实验结果与分析

本文将第3节抓取的“批评”事件的语料作为训练集。使用同样的语料抓取方法对给出的另一个核心事件：某人“打”某人，从Web抓取了短文本共计42827个，经过同样的预处理过程，得到了这个事件的事件语料。把某人“打”某人事件的语料作为测试集。训练集、测试集都做了相关事件和角色联系的人工标记。在此基础上，对相关事件挖掘部分和角色联系发现部分，分别给了评价指标和试验结果的相关分析。

1)用两个指标来评价相关事件挖掘。相关事件挖掘准确率：

$$EP=\frac{E\text{中事件被人工判定为与}e_0\text{相关的事件对应的事件短语总数}}{E\text{中全部事件对应的事件短语总数}} \tag{14}$$

相关事件挖掘召回率：

$$ER=\frac{E\text{中全部事件对应的事件短语总数}}{D\text{中人工标记的与}e_0\text{相关的事件短语总数}} \tag{15}$$

在“批评”事件的语料中经过多次调节，最终将挖掘算法中二元词语外部、内部点互信息预设阈值 $MI_{外}$、$MI_{内}$ 分别设定为：$MI_{外}=0.15$，$MI_{内}=0.1$。

在测试集上获得的试验结果如表4所列。

表4 测试集的相关事件挖掘结果

	二元词语	三个词扩展	四个词及以上扩展	总计
EP	88.1%	73.3%	54.4%	73.0%
ER	83.6%	70.8%	46.7%	68.2%

从表4中可以看出算法对二元词语挖掘取得了比较好的

结果，准确率、召回率都取得了最高值。这些短语大多是动宾、动补、主谓短语，其含义明确，语义也较完整。两个词以上的事件短语的语义更加复杂，算法的效果有所降低。其主要原因是：1)被错误召回的事件短语主要是普通常用的高频短语，这类短语与核心事件没有很好地关联；2)包含的词数越多，事件短语数越少，给识别和挖掘造成较大困难，也导致了准确率、召回率降低；3)句法树扩展规则还未能完全覆盖所有的模式，特别是事件降级为事元[14]，形成事件嵌套的句子还没有涉及到，导致不能把较复杂的相关事件挖掘出来。

2)用两个指标来评价角色联系发现。角色联系发现准确率：

$$RP=\frac{E_w\text{ 中角色标记正确的总数}}{E_w\text{ 中事件总数}} \quad (16)$$

角色联系发现召回率：

$$RR=\frac{E_w\text{ 中事件总数}}{E\text{ 中人工标记为与 }e_0\text{ 有角色联系的事件总数}} \quad (17)$$

在角色联系发现算法中，δ_{kx}，δ_{ky}大于预设阈值δ_0时，就将该事件标记为对应的角色。可见预设阈值δ_0对角色联系发现有很大影响，在训练集中δ_0对准确率、召回率的影响曲线如图3所示，图中横坐标表示预设阈值。本文最后取阈值为0.7，以获得较好的准确率 。

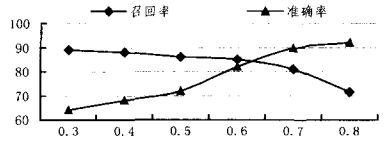


图3 训练集中δ_0对角色联系发现评测的影响曲线

预设好参数后，把人工标注的相关事件集合作为输入的相关事件集合。在测试集上进行测试，获得了81.4%的准确率和68.5%的召回率。测试集准确率有所降低，召回率有较大的损失。虽然角色转换规则还很不完善，针对一个事件短语的角色标记还不很准确，但其多次的标记却有很高的可信度。事件短语的频率对角色联系发现的结果影响较大，当频率较高时比较容易发现其角色联系。在被错误判定的角色联系的事件中，一部分事件语义本身不完整，而一部分事件则本身在两个角色人中没有明显的倾向，或者两种角色倾向都符合事件的语义逻辑。

7 方法的创新和适用

本文提出的事件多元组即反映了事件本身，还记录了事件间的联系；事件多元组适应各类事件，有较好的通用性。本文提出了二元词语的概念。二元词语本身含有词汇依存信息，可以看成是组成依存句法树的基本部件。同时，二元词语在唯一的一个词汇依存关系下，构成二元词语的两个词的语义也做到了基本固定。也就是说，二元词语中的每个词的词义通过两个词的词汇依存关系已经基本确定下来了。我们认为二元词语是一个语义和表述形式的综合体，既承载了确定的语义信息，又为表达复杂语义做了形式化准备。特别是我们可以根据不同的需要，结合词义添加新的二元词语词汇依存关系，使二元词语的语义更细致。

本文给出的已知事件中，明确区分了事件的不同角色。这种区分也体现在核心事件的事件短语的构造过程中，不同的事件角色在句式或句模里的位置是明晰的。通过我们提出的方法在事件语料中找到组成相关事件的二元词语，在扩展规则的限制下既获取了相关事件的表述形式，又能得到相关事件的内容。当然，对于给定的核心事件，其相关事件以哪些句式或句模表述，又应该添加哪些语义更细致的二元词语词汇依存关系是影响挖掘效果的关键因素。但这并不能妨碍二元词语可以适用于各类事件的挖掘。

本文的事件角色联系发现，主要讨论了事件的角色人。在事件相关的前提下，此方法可以获得一个事件中的角色人又参与了哪些相关事件，从而在一定程度上克服了代词指代消歧这一难题。

结束语 本文试图围绕某一个核心事件，挖掘与之相关的事件，希望找到一个事件发生时的前后事件。然后，继续发现这些相关事件与核心事件间的逻辑联系，即角色联系。本文提出了基于二元词语来挖掘相关事件的方法，通过构造句法扩展树来标出事件短语；还提出了事件角色联系的发现方法，这个方法恰恰是利用了代词的作用，使事件通过角色人得以关联。实验结果显示，本文挖掘到大量相关事件和事件间的角色联系。相关事件挖掘算法的准确率和召回率分别达到73.0%和68.2%；角色联系发现算法的准确率和召回率分别达到81.4%和68.5%。

下一步的研究工作，将集中在二元词语的词汇依存关系的添加与扩充，探究词汇依存与语义表达的内在联系；还要对句法扩展规则作进一步完善。

参 考 文 献

[1] 曹存根，丰强泽，等. Progress in the Development of National Knowledge Infrastructure[J]. 计算机科学技术学报(英文版)，2002，17(5)：523-534

[2] Hearst M A. Text data mining：Issues，techniques，and the relationship to information access[C]// Presentation notes for UW/MS workshop on data mining. 1997

[3] Nelson K，Gruendel J. Event knowledge：structure and function in development[M]. Baker & Taylor Books，1986

[4] Zhou W，Liu Z，Zhao Y，Xu L，Qiang Y. A Semi-automatic Ontology Learning Based on WordNet and Event-based Natural Language Processing Technologies[C]// Proceedings of ICIA'06 Conference (Accepted and Registered). 2006

[5] 吴平博，陈群秀，马亮. 基于事件框架的事件相关文档的智能检索研究[J]. 中文信息学报，2003，7：25-30

[6] 梁晗，陈群秀，吴平博. 基于事件框架的信息抽取系统[J]. 中文信息学报，2006，20：40-46

[7] 姜吉发. 一种跨语句汉语事件信息抽取方法[J]. 计算机工程，2005，31(2)：27-29

[8] Talmy L. Toward a Cognitive Semantics[M]. Cambridge：MIT Press，Vol. 1，2000

[9] 韦卓，赵克，易帅. 面向领域的动词事件聚类[J]. 计算机工程与科学，2008，30(3)：133-135

[10] 吴中兴，赵克，胡钢伟，等. 概念从属树——一种新的树模型设计[J]. 计算机应用，2004，24(z1)：99-100

[11] NIST. Topic Detection and Tracking Evaluation[OL]. http://www.nist.gov/speech/tests/tdt/index.htm，2007

[12] Nallapati R，Ao Feng，Fuchun Peng. Event threading within news topics[C]// Proc. of the 13th ACM Conf on Information and Knowledge Management. New York：ACM，2004：446-453

[13] Liu Ting，Ma Jinshan，Li Sheng. Building a Dependency Treebank for Improving Chinese Parser[J]. Journal of Chinese Language and Computing，2006，16(4)：207-224

[14] 鲁川，王玉菊. 汉字信息语法学(第一版)[M]. 山东：山东教育出版社，2008：239-242，308-310

[15] Salton. Term weighting approaches in automatic text retrieval [J]. Information processing and management，1988，24(5)：513

[16] 马金山. 基于统计方法的汉语依存句法[D]. 哈尔滨：哈尔滨工业大学，2007