基于标签搭配与迭代融合的原子事件抽取方法

张贺

摘　要: 为完备用于原子事件抽取的迭代融合方法所需的标签搭配规则并验证其有效性,分析了已有事件抽取技术、N 元模型理论和标签搭配统计方法,并结合中文特点,提出了一种标签搭配规则验证与自动发现方法,并通过迭代融合方法将标签搭配规则应用于原子事件抽取.该方法包括标签搭配统计、标签搭配筛选、标签搭配迭代融合以及原子事件抽取等步骤.实验结果表明该方法能够有效地验证预定义标签搭配规则并发现新的标签搭配规则,同时,标签搭配迭代融合方法对原子事件抽取是帮助的.

关键词: 标签搭配;迭代融合;原子事件抽取;N 元模型

1　引言

事件抽取隶属于信息抽取领域,主要研究如何把自然语言形式表达的事件以结构化的形式呈现出来,如什么人、在什么地方、什么时间、做了什么事等.随着互联网的发展,事件抽取成为自然语言处理领域的研究热点.目前,事件抽取技术在文本蕴含[1],社交网络分析[2],信息检索[3],股价预测[4],指代消解[5],医疗系统[6]等领域都取得了较好的应用效果.

目前事件抽取的研究大都是针对 MUC、TDT、ACE、TAC 等评测任务或在此基础上展开的[7-15].传统事件抽取方法是将事件识别作为事件分类问题,通过机器学习方法或者基于事件类型模版填充方法来发掘文本中的事件,这类方法的通用性并不好,往往只针对某个领域的事件、某个类型的事件或某个主题的事件进行抽取,移植性并不理想.此外,如果要对段落、篇章、文档集进行全面透彻地分析,不仅要抽取文本中的主线事件,还要抽取其中的附属事件,在事件类型约束前提下,是很难做到这一点的.目前,与类型无关的事件抽取相关研究并不多,本文旨在提出一种事件类型无关的事件抽取方法,该方法不受事件类型的约束,能够利用统计和规则相结合的方法抽取一段文本中的原子事件.

事件抽取方法可以分为统计方法和规则方法两大类.Boros 等[7]通过有限的先验领域知识训练得到的神经网络模型来抽取特定领域的事件角色;传统的 ACE 事件抽取任务包含事件触发词抽取和事件论元抽取等步骤,Li 等[8]利用谓词结构理论将事件触发词抽取模型和事件论元抽取模型进行联合以避免错误传播,同时采用大量的全局特征抽取触发词与论元之间的依赖关系;李培峰等[9]利用触发词以及触发词上下文一致性的组合语义特征解决了中文事件抽取的触发词识别任务中未登录触发词的识别和分词错误导致的触发词无法识别两大问题;丁效等[10]通过首先对音乐领域事件词进行聚类,然后采用关键词与触发词相结合的方法识别事件类型,最后采用了基于最大熵的方法进行事件元素识别;肖升等[11]提出了一种基于超图的事件类型识别方法,在一定程度上避免向量空间模型的独立性假设影响事件类型识别;Llorens 等[12]通过 CRF 模型进行语义角色标注,并应用于 TimeML 的事件抽取,提升了系统性能.

N 元模型和标签搭配是比较成熟的技术,近年应用依然十分广泛.Yeh 等[16]提出一个基于 N 元模型的倒排索引方法来解决中文拼写错误检查和纠正的问题; Noji 等[17]尝试将 N 元模型的思想用于话题语言模型,对 Wallach 提出的贝叶斯话题语言模型进行了改进;车万翔等[18]在识别可靠依存关系弧的问题时,利用 N 元模型和标签搭配作为逻辑回归模型的重要特征进行模型训练;唐都钰等[19]提出一个使用单词嵌入的 Twitter 情感分类方法,将文本中的情绪信息编码到一串连续的词向量中,然后采用 N 元模型和三层神经网络模型进行学习.单煜翔等[20]通过扩展 N 元模型简化了大词汇量连续语音识别器中解码器的实现,提升了语言模型预测速度,使高阶语言模型预测成为可能;龚正仙等[21]提出基于 N 元时态模型的统计机器翻译方法,对已有的基于短语的统计机器翻译系统进行了改进.

ACE、TAC 等评测任务对事件类型进行了限定,包含生活、移动、商业、战争等 8 大类事件.受 N 元模型和标签搭配思想启发,本文的原子事件抽取旨在对整篇文档进行事件类型无关的事件抽取,即将文档中所有符合原子事件结构的信息抽取出来,实现对篇章整体的结构化而不仅仅是局部信息的结构化.在标签搭配方法中,

所用标签搭配规则的质量对标签搭配迭代融合方法的效果起着关键性作用.本文在 N 元模型的基础上,利用统计与规则相结合的方法实现对预定义标签搭配规则的验证以及新标签搭配规则的发现,首先利用 N 元标签搭配规则统计模型进行统计,然后通过统计筛选方法和语言筛选规则筛选出最终标签搭配规则,最后基于标签搭配规则利用迭代融合方法进行原子事件抽取.使用标签搭配的方法进行原子事件抽取不需要对文本进行句法分析、语义分析,直接在词法分析的基础之上根据标签搭配规则进行原子事件成分融合,实验结果表明该方法是有效的.

2 系统结构

在基于标签搭配与迭代融合的原子事件抽取方法中,首先通过标签搭配规则验证与自动发现方法获得最终标签搭配规则;接着就可以基于最终标签搭配规则利用迭代融合进行原子事件抽取,方法的框架图如图 1 所示.

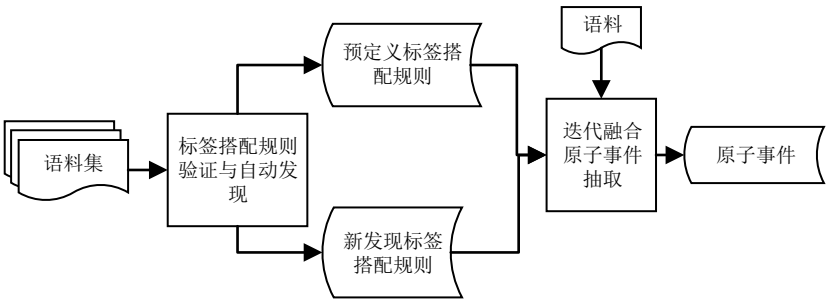

Fig.1 Model overview

图 1 模型框架图

标签搭配规则验证与自动发现方法是以 N 元模型为基础的,经过语料库预处理、N 元标签搭配统计模型、标签搭配统计筛选、标签搭配语言规则筛选、预定义标签搭配规则验证以及新标签搭配规则发现等六个阶段的处理后,得到最终标签搭配规则.图 2 为基于 N 元模型的标签搭配规则验证与自动发现方法的流程图.

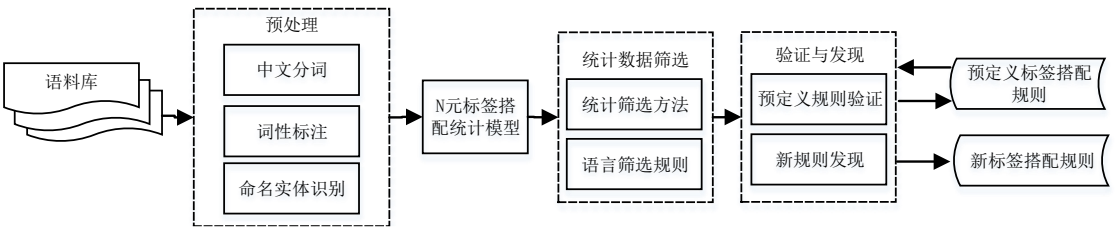

Fig.2 N-gram model based tag collocation rules verification and automatic discovery method

图 2　基于 N 元模型的标签搭配规则验证与自动发现方法

在语料库预处理阶段,首先利用斯坦福自然语言处理工具(http://nlp.stanford.edu/software/index.shtml)对语料进行中文分词、词性标注和命名实体识别.预处理之后的语料包含滨州中文树库标签集合[22]中除 IJ、ON、PU、FW 之外的 29 种词性标签,同时还包含 3 种命名实体标签,即 PERSON(人名实体)、GPE(地缘政治实体)和 LOC(地点实体).

在标注了词性和命名实体的语料基础上,利用标签搭配统计模型统计其中的二元标签搭配数据以生成二元标签搭配数据矩阵.根据二元标签搭配数据矩阵计算得出相应的条件概率矩阵,利用统计方法从中筛选出达到标签搭配可信度的标签搭配规则.由于统计方法自身的局限性,此时的标签搭配规则存在一定的不足,本文结合自然语言处理的方法,进一步分析统计筛选方法筛选结果,总结出一些特征规律并将其表示成语言筛选规则.在统计筛选方法筛选结果的基础上,利用语言筛选规则筛选出符合语言规则的标签搭配规则,得到最终的二元标签搭配规则.

将最终二元标签搭配规则与预定义二元标签搭配规则进行对照,对预定义二元标签搭配规则进行验证,最

终二元标签搭配规则与有效预定义二元标签搭配规则的差集即为新发现二元标签搭配规则.通过二元标签搭配的多次迭代融合,能够将文本中多个临近标签进行融合,为了减少迭代次数,可以采用一系列三元标签搭配规则,例如“NN CC NN”(NN 为名词标签,CC 为联结词标签)、“P NN LC”(P 为介词标签,LC 为方位词标签)等在语用上经常进行搭配出现的三元标签搭配,通过对大量文本实例分析总结,预定义标签搭配规则中包含了一系列三元标签搭配规则.

标签搭配规则验证与自动发现方法得到的最终标签搭配规则,是原子事件抽取方法的重要输入.在原子事件抽取方法中,首先对源文档进行分词,然后在分词结果的基础上进行词性标注和命名实体识别,之后根据最终标签搭配规则将预处理后的文本进行迭代融合.将迭代融合之后的文本进行句子切分,对不同结构的原子句采用相应的方法进行原子事件成分的抽取.算法细节会在后面的章节进行详细叙述.

3 N 元标签搭配统计模型

N 元模型的基本思想:一个元素的出现与其上文环境中出现的元素序列密切相关,第 n 个元素的出现只与前 $n-1$ 个元素相关,而与其他任何元素都不相关,设 $e_1e_2...e_n$ 是长度为 n 的元素串,则元素序列的出现概率如公式(1)所示:

$$P_n = P(e_n \mid e_1, e_2, ..., e_{n-1}) \tag{1}$$

当 N 取值为 1 时称为一元模型,即元素上下文无关;N 取值为 2 时称为二元模型,依此类推可有三元模型、四元模型等.

从公式(1)可以看出,为了计算元素 e_n 的出现概率,需要计算它前面所有元素的出现概率.如果 N 的取值较大,就需要十分庞大的计算量,而这是十分影响模型效率的.针对本文提出的方法而言,一元模型未考虑标签与标签之间的上下文关系,因此无法体现出标签与标签之间的搭配关系;通过分析前人的工作成果可知,在利用 N 元模型进行分类任务时,N>3 时性能变化不大,甚至有所降低,综合以上因素,本文主要采用二元模型作为标签搭配统计模型,通过二元标签搭配的迭代融合可以扩展到多元,实现多元标签搭配融合.

3.1 标签搭配统计模型

将一篇预处理之后的语料表示为“词-标签”对序列“$w_1/t_1\ w_2/t_2\ w_3/t_3...w_n/t_n$”,其中 w_i 表示词,t_i 表示 w_i 对应的词性标签或命名实体标签,“w_i/t_i”表示一个“词-标签”对,n 表示以词为单位计算的语料预处理结果的长度.

二元标签搭配统计模型设计如下:初始化一个大小为 N*N,全部元素为 0 的二元标签搭配数据矩阵(BTM),N 为标签种类数目,本文中标签种类包含 29 种词性标签和 3 种命名实体标签,故 N 为 32,BTM 用公式(2)表示:

$$BTM = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix} \tag{2}$$

BTM 中的元素 c_{ij}(1<=i<=N,1<=j<=N)表示第 i 种标签在前和第 j 种标签在后的标签搭配出现的次数.对每篇语料对应的“词-标签”对序列,从“词-标签”对序列的第一个“词-标签”对开始,依次遍历相邻两个“词-标签”对“$w_j/t_j\ w_{j+1}/t_{j+1}$”(1<=j<=n-1)的标签搭配“$t_j\ t_{j+1}$”,并根据 t_j 的标签种类 p 和 t_{j+1} 的种类 q 将 BTM 中的元素 c_{pq} 加 1,统计完语料库中全部语料为止.在二元标签搭配“$t_j\ t_{j+1}$”中,本文称第一个标签“t_j”为首标签,第二个标签“t_{j+1}”记为尾标签.

3.2 统计筛选方法

对于二元标签搭配规则“$t_p\ t_q$”,计算 t_p 为首标签时,尾标签是 t_q 的条件概率,称之为“$t_p\ t_q$”的首概率,记作 $HP(t_p\ t_q)$,用公式(3)表示:

$$HP(t_p t_q) = c_{pq} \Big/ \sum_{j=1}^{N} c_{pj} \tag{3}$$

对于二元标签搭配规则“t_p t_q”,还要计算 t_q 作为尾标签时,首标签是 t_p 的条件概率,称之为“t_p t_q”的尾概率,记作 $TP(t_p\ t_q)$,用公式(4)表示:

$$TP(t_p t_q) = c_{pq} \Big/ \sum_{j=1}^{N} c_{jq} \tag{4}$$

公式(3)和(4)中 c_{qp} 表示 *BTM* 中第 p 行与第 q 列的数据,N 表示 *BTM* 中标签种类数目.根据 *BTM*、公式(3)和(4),能够计算出首概率矩阵(*HPM*)和尾概率矩阵(*TPM*),分别用公式(5)与(6)表示:

$$HPM = \begin{bmatrix} HP(t_1t_1) & HP(t_1t_2) & \dots & HP(t_1t_N) \\ HP(t_2t_1) & HP(t_2t_2) & \dots & HP(t_2t_N) \\ \dots & \dots & \dots & \dots \\ HP(t_Nt_1) & HP(t_Nt_2) & \dots & HP(t_Nt_n) \end{bmatrix} \tag{5}$$

$$TPM = \begin{bmatrix} TP(t_1t_1) & TP(t_1t_2) & \dots & TP(t_1t_N) \\ TP(t_2t_1) & TP(t_2t_2) & \dots & TP(t_2t_N) \\ \dots & \dots & \dots & \dots \\ TP(t_Nt_1) & TP(t_Nt_2) & \dots & TP(t_Nt_n) \end{bmatrix} \tag{6}$$

根据 *HPM* 和 *TPM*,通过标签搭配可信度来进行初步的筛选,综合考虑标签搭配的完备性与有效性,本文将标签搭配可信度设置为均值 100%/N=3.125%,其中 N 表示 *HPM* 和 *TPM* 中标签种类数目.

对经过标签搭配可信度筛选的标签搭配规则进行分析,其中大部分标签搭配规则是符合汉语使用特点的,但仍存在不好的搭配规则.其原因包括:(1)统计方法没有考虑标签在使用过程中的特性,而是把所有的标签都统一对待.(2)选取的语料库与选用的标签搭配可信度并不是完美相容,无法最大发挥统计方法的作用.(3)汉语是一种灵活的语言,其中的标签搭配规则自然也是灵活多变,但是高可信度上的标签搭配规则在语言规则上不一定合理.因此,需要通过增加语言知识来进一步筛选.

3.3 语言筛选规则

在二元标签搭配统计筛选方法筛选结果的基础上,根据语言规则,利用标签的语义信息,进一步筛选符合语言规则的二元标签搭配规则.通过对本方法所使用标签体系中标签的语义与二元标签搭配统计数据进行分析,发现标签 P 仅作为二元标签搭配的首标签时符合语义;尽管部分标签经常出现在标签搭配起始位置,但是由于 CC、DEG 等具有联结功能的标签存在,这些标签也可以作为标签搭配的尾标签;ETC、LC 等标签仅作为二元标签搭配的尾标签时符合语义;CC、DEG、NN 等标签作为二元标签搭配的首标签和尾标签皆符合语义;AS、BA、PU 等标签是标签搭配分隔符或非标签搭配成分,因此不作为标签搭配的成分;动词是一比较特殊的词类,基于标签搭配迭代融合方法的原子事件抽取是以原子事件谓词驱动的,而一般情况下原子事件谓词就是单个动词,不需要对动词进行过多的融合,故要筛除全部与动词相关的二元标签搭配规则.具体二元标签搭配语言筛选规则如表 1 所示.

Table 1 Bi-gram tag collocation filtering based on linguistic rules

表 1 二元标签搭配语言筛选规则

标签类型	标签
仅作首标签	P
仅作尾标签	ETC,LC
作首尾标签均可	AD,DT,JJ,VA,CC,CD,DEC,DEG,DEV,MSP,NN,NR,OD,PN,PERSON,GPE,LOC,NT
非搭配标签成分	AS,BA,PU,CS,LB, ON,SB,SP,MSP

4 原子事件抽取算法

得到最终标签搭配规则之后,就可以进行原子事件抽取,在原子事件抽取算法中,首先要基于标签搭配规则对预处理后的文本进行迭代融合,当文本中存在能够匹配某一标签搭配规则时,就将相应的标签搭配进行融合,融合结果分为三类:(1)相同标签的融合结果不变;(2)第二个标签为“DEG”或“DEC”则融合结果为“JJ”;(3)其他情况下,融合结果为第二个标签.以句子“思科公司是全球最大的互联网设备供应商. ”为例,图 3 展示了这个句子迭代融合的过程,其中方框标识的部分为下一轮迭代需要融合的成分,在第四轮迭代之后,句子中没有需要迭代融合的成分,迭代过程终止.

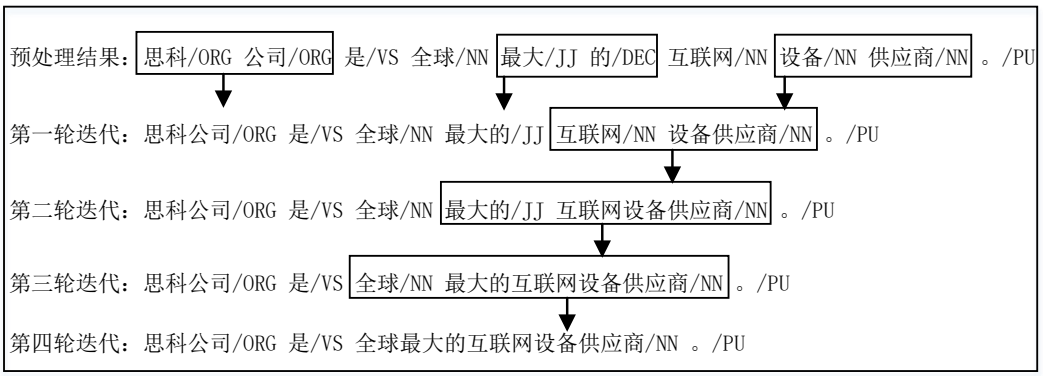

Fig.3 Tag iterative fusion

图 3 标签迭代融合过程

在进行原子事件抽取时,本文采用基于原子句句子结构的方法,因此要将文本中的句子根据逗号、嵌套关系和并列关系切分为原子句.一般情况下,一个句子中会存在多个原子事件,其中以逗号分割居多,还存在一些含有嵌套关系和并列关系的原子事件.嵌套关系根据嵌套关系动词表和原子句中原子事件谓词个数进行识别,并列关系根据原子句中原子事件谓词个数进行识别,当原子句中同时符合嵌套关系和并列关系的条件时,嵌套关系的优先级高于并列关系.以预处理后的句子“他们/PN 估计/VV 他的腿/NN 被/SB 摔伤/VV ,/PU 他/PN 不能继续游泳/VV 而/MSP 被/SB 水流/NN 卷走/VV ./PU”为例.根据逗号可以切分为下面两个原子句:

(1)他们/PN 估计/VV 他的腿/NN 被/SB 摔伤/VV;

(2)他/PN 不能继续游泳/VV 而/MSP 被/SB 水流/NN 卷走/VV;

在原子句(1)中的“估计”一词存在于嵌套关系动词表并且存在两个原子事件谓词,故对原子句(1)进行切分.在切分时,首先生成一个新的原子句,内容为嵌套关系动词之后的成分,之后将原句中嵌套关系动词之后的成分融合为名词成分.原子句(1)的切分结果为:

(1-1)他们/PN 估计/VV 他的腿被摔伤/NN;

(1-2)他的腿/NN 被/SB 摔伤/VV;

在原子句(2)中,存在两个原子事件谓词“不能继续游泳”和“卷走”,故对原子句(2)进行切分.在切分时要注意判断原子事件边界,先找到原子事件谓词,然后向两边扩展找到最近的名词成分或者原子事件谓词.原子句(2)的切分结果为:

(2-1)他/PN 不能继续游泳/VV;

(2-2)被/SB 水流/NN 卷走/VV;

本文将原子句的基本句子结构分为三类,认为文本中句子均可以由这三类基本句子结构扩展联合得出,这三类基本句子结构即一般句式、“把”字句式和“被”字句式,其中一般句式的基本结构为“主体+事件谓词+客体”,“把”字句式的基本结构为“主体+把+客体+事件谓词”, “被”字句式的基本结构为“客体+被+主体+事件谓词”.以原子句(1)为例,它可由原子句(1-1)和原子句(1-2)联合而成,其中原子句(1-1)的句子结构属于一般句式,原子句(1-2)的句子结构属于“被”字句式.

在抽取时,由于原子句中谓词是唯一的,因此每个原子句中的谓词即为原子事件谓词.而在抽取原子事件主

体与原子事件客体时,需要考虑原子句句子结构,如果是一般句式,以原子事件谓词为起点,向左搜索到的第一个名词性成分(普通名词、专有名词、人称代词、命名实体等)为原子事件主体,向右搜索到的第一个名词性成分为原子事件客体;如果是“把”字句式,以介词“把”或“将”为起点,向左搜到的第一个名词性成分为原子事件主体,向右搜索到的第一个名词性成分为原子事件客体;如果是“被”字句式,以介词“被”为起点,向右搜到的第一个名词性成分为原子事件主体,向左搜索到的第一个名词性成分为原子事件客体.算法1描述了原子事件抽取算法流程,其中(ht_i,tt_i,ct_i)表示一条二元标签搭配规则,ht_i和tt_i分别表示标签搭配规则的首标签和尾标签,ct_i表示ht_i与tt_i融合的结果标签.在迭代融合过程中,三元标签搭配融合方法与二元标签搭配融合方法类似,不再赘述.

Algorithm 1 Atomic event extraction algorithm

算法1 原子事件抽取算法

名称:原子事件抽取算法

输入:经过预处理的文本T={w_1/t_1,w_2/t_2,...,w_n/t_n,...};

标签搭配规则集合TC={(ht_1,tt_1,ct_1),(ht_2,tt_2,ct_2),...,(ht_n,tt_n,ct_n),...};

输出:原子事件集合E

步骤 1:**for** w_i/t_i w_{i+1}/t_{i+1} **in** T **do**:

步骤 2: **for** ht_j tt_j ct_j **in** TC **do**:

步骤 3: **if** t_i==ht_j **and** t_{i+1}==tt_j **do**:

步骤 4: 将T中w_i/t_i w_{i+1}/t_{i+1}融合为w_iw_{i+1}/ct_j;

步骤 5: **end if**;

步骤 6: **end for**;

步骤 7:**end for**;

步骤 8:E=Φ;

步骤 9:根据点号将T切分为句子集合S={s_1,s_2,...,s_n,...};

步骤 10:**for** s_i **in** S **do**:

步骤 11: 将s_i切分为原子句集合AS={as_1,as_2,...,as_n,...};

步骤 12: **for** as_j ***i*n** AS **do**:

步骤 13: 判断as_j句子结构ST;

步骤 14: 抽取原子事件谓词AEP;

步骤 15: **if** ∃AEP **do**:

步骤 16: 根据ST,抽取原子事件主体AA和原子事件客体AP;

步骤 17: **if** ∃(AA∨AP) **do**:

步骤 18: 原子事件e={AEP,AA,AP};

步骤 19: E=E∪e;

步骤 20: **end if**;

步骤 21: **end if**;

步骤 22: **end for**;

步骤 23:**end for**;

5 实验结果分析

5.1 标签搭配规则验证与自动发现

本文对 NTCIR-8 中使用的新华日报新闻语料库(500MB)进行了实验,该语料库包含新华日报 2002 年至 2005 年全部的新闻,共计新闻语料 262906 篇,包含二元标签搭配 58636844 对.

对语料库进行统计之后得到的 *BTM* 涵盖了 1024 项二元标签搭配.根据标签搭配可信度从 *HPM* 中筛选出标签搭配规则 188 项,从 *TPM* 中筛选出标签搭配规则 188 项.经统计,在 *HPM* 筛选结果和 *TPM* 筛选结果中均出

现的二元标签搭配规则有 92 项,只在 *HPM* 筛选结果中出现的有 96 项,只在 *TPM* 筛选结果中出现的有 96 项,故统计方法筛选的二元标签搭配规则包含 284 项.

根据语言规则对统计方法筛选所得 284 项标签搭配规则进一步筛选,最终去掉了 134 项,保留了 150 项.在去掉的 134 项标签搭配规则中,不少标签搭配规则的首概率或尾概率都具有较高的可信度,但是在语言规则上不宜进行搭配,例如标签搭配“LB NN”的首概率为 60.24%,“NN LB”的首概率为 27.95%,也就是说“被”的前后出现名词的概率都比较高,这是符合语言学规律的,但是标签搭配规则的目标是原子事件抽取,如果将“LB”和“NN”进行合并会导致无法完整地抽取原子事件成分,因此,在统计方法筛选标签搭配之后,再用语言规则筛选是十分有必要的.保留的 150 项标签搭配规则即最终二元标签搭配规则,其中在 *HPM* 筛选结果和 *TPM* 筛选结果中均出现的有 65 项;只在 *HPM* 筛选结果中出现的有项 32 项;只在 *TPM* 筛选结果中出现的有 53 项.标签搭配数据筛选结果如表 2 所示.

Table 2 Tag collocation filtering result

表 2 标签搭配数据筛选结果

筛选方法	标签搭配出现位置	*HPM* 和 *TPM* 筛选结果中均出现	仅在 *HPM* 筛选结果中出现	仅在 *TPM* 筛选结果中出现	共计
统计筛选方法	搭配项数	92	96	96	284
	百分比	32.40%	33.80%	33.80%	100%
语言筛选规则	搭配项数	65	53	32	150
	百分比	43.33%	35.33%	21.33%	100%

关于预定义二元标签搭配规则的验证方法,本文通过准确率(*Precision*)、召回率(*Recall*)和 *F1-measure* 来进行度量,其具体计算方法如公式(7)、(8)和(9):

$$\mathrm{Pr}\,ecision = \frac{\text{合格预定义二元标签搭配规则数目}}{\text{预定义二元标签搭配规则数目}} \tag{7}$$

$$Rcall = \frac{\text{合格预定义二元标签搭配规则数目}}{\text{最终二元标签搭配规则数目}} \tag{8}$$

$$F1\text{-}measure = \frac{2\times \mathrm{Pr}\,ecision \times \mathrm{Re}\,call}{\mathrm{Pr}\,ecision + \mathrm{Re}\,call} \tag{9}$$

预定义二元标签搭配规则 82 项,能够与最终二元标签搭配规则进行匹配的有 47 项.经统计,预定义二元标签搭配规则验证结果如表 3 所示.

Table 3 Predefined bi-gram tag collocation rules verification result

表 3 预定义二元标签搭配规则验证结果

Precision	Recall	F1-measure
57.31%	31.30%	40.49%

最终二元标签搭配规则与合格预定义二元标签搭配规则的差集中所含的 103 项为新发现二元标签搭配规则,占最终二元标签搭配规则的 68.7%.

通过表 3 中的评测数据以及新发现的 103 项二元标签搭配规则可以看出,经过本方法对预定义二元标签搭配规则的验证以及对新标签搭配规则的发现,去掉了 35 项可信度较低的预定义二元标签搭配规则,发现了 103 项新的二元标签搭配规则,极大的提升了二元标签搭配规则的合理性.

5.2 原子事件抽取

虽然已有大量关于事件抽取的相关评测会议与相应的评测语料,但是这类会议主要关注特定类型事件模版填充以及事件要素识别[13-15],其中 ACE 中抽取 33 种类型事件的最高 F 值为 53.9%[13].目前,国内外关于事件类型无关的事件抽取研究较少,且尚未发现可供公开评测的语料库.本文采用实验语料来自于 NTCIR-9 RITE 任务,包含 1414 个句子.我们对该语料进行原子事件标注,语料集中包含 2701 个原子事件,2116 个主体成分,1999 个客体成分,2651 个原子事件谓词.标注过程如下:

（1）将标注人员分为 3 组,分别独立地标注,标注过程中仅允许组内讨论;

（2）对三组标注后的语料进行一致性检查,认为标注一致的原子事件成分是正确的;

（3）对于标注不一致的原子事件成分,3 组标注人员进行投票表决.

将合格预定义规则与新发现规则分为:预定义二元标签搭配规则、预定义三元标签搭配规则和新发现标签搭配规则.将三种标签搭配规则进行组合,得到 8 种不同的标签搭配规则组合,然后利用 8 种不同的标签搭配规则组合分别对实验语料进行原子事件抽取.标签搭配规则组合如表 4 所示.

Table 4 Tag collocation rule combinations

表 4 标签搭配规则组合

方法	方法描述
None	仅使用无标签搭配规则
PreBi	仅使用预定义二元标签搭配规则
PreTri	仅使用预定义三元标签搭配规则
DisBi	仅使用新发现二元标签搭配规则
PreBi+PreTri	同时使用 PreBi 和 PreTri
PreBi+DisBi	同时使用 PreBi 和 DisBi
PreTri+DisBi	同时使用 PreTri 和 DisBi
All	同时使用 PreBi、PreTri 和 DisBi

关于原子事件成分(原子事件主体、原子事件客体、原子事件)抽取结果的评测方法,本文通过准确率(*Precision*)、召回率(*Recall*)和 *F1-measure* 来进行度量,其具体计算方法如公式(10)、(11)和(12):

$$\mathrm{Pr}ecision = \frac{\text{正确抽取原子事件成分数目}}{\text{自动抽取原子事件成分数目}} \tag{10}$$

$$\mathrm{Re}call = \frac{\text{正确抽取原子事件成分数目}}{\text{人工标注原子事件成分数目}} \tag{11}$$

$$F1\text{-}measure = \frac{2*\mathrm{Pr}ecision*\mathrm{Re}call}{\mathrm{Pr}ecision+\mathrm{Re}call} \tag{12}$$

原子事件主体成分抽取结果和原子事件客体成分抽取结果分别如表 5、表 6 所示:

Table 5 Atomic event extraction results of subjective elements

表 5 原子事件主体成分抽取结果

方法	Precision	Recall	F1-measure
None	40.11%	42.86%	41.44%
PreBi	49.95%	51.13%	50.54%
PreTri	41.30%	43.95%	42.58%
DisBi	49.38%	49.15%	49.27%
PreBi+PreTri	50.85%	52.03%	51.44%
PreBi+DisBi	53.25%	52.65%	52.95%
PreTri+DisBi	48.67%	49.10%	48.88%
All	53.32%	53.17%	53.24%

Table 6 Atomic event extraction results of objective elements

表 6 原子事件客体成分抽取结果

方法	Precision	Recall	F1-measure
None	29.22%	30.47%	29.83%
PreBi	41.15%	42.47%	41.80%
PreTri	32.85%	34.17%	33.50%
DisBi	42.77%	41.87%	42.32%
PreBi+PreTri	43.23%	44.57%	43.89%
PreBi+DisBi	44.08%	43.62%	43.85%

PreTri+DisBi	43.16%	43.07%	43.11%
All	44.19%	44.17%	44.18%

通过表 5 和表 6 中的数据进行对比,原子事件客体成分的抽取效果略低于原子事件主体成分抽取效果,究其原因,在汉语文本中描述一个事件时,用于对原子事件客体进行补充说明的附加成分较多,而用于对原子事件主体进行补充说明的附加成分相对较少,因此在采用相同方法的前提下,原子事件客体成分的抽取效果比原子事件主体成分抽取效果是要稍低的.

从表 5 可以看出,在抽取原子事件主体成分时,同时采用预定义二元标签搭配规则、预定义三元标签搭配规则和新发现规则的“方法 All”效果最好,准确率为 53.32%,召回率为 53.17%,F 值为 53.24%;而未采用标签搭配规则的“方法 None”效果最差,准确率、召回率和 F 值均比“方法 All”低 10%以上,比采用预定义二元标签搭配规则和预定义三元标签搭配规则的“方法 PreBi+PreTri”低 10%左右,因此,通过采用标签搭配规则进行原子事件主体成分抽取是有效的.此外,“方法 PreBi+PreTri”和“方法 All”的差别在于是否采用了新发现标签搭配规则,从实验结果来看,在标签搭配规则中添加了新标签搭配规则之后,抽取效果也有所提升.表 6 中抽取原子事件客体成分时采用不同方法取得的数据所呈现的趋势与表 5 相似, “方法 None”效果最差,在预定义标签搭配的基础上添加新发现标签搭配之后效果都有所提升,而同时使用了预定义标签搭配规则和新发现标签搭配规则的“方法 All”的效果最好,其 F 值为 44.18%.

在评价原子事件整体抽取效果时,除了兼顾原子事件主体成分和原子事件客体成分的抽取是否准确外,还要考虑原子事件谓词的抽取是否准确.利用前文所述原子事件抽取算法进行原子事件抽取的结果如表 7 所示:

Table 7 Atomic event extraction results

表 7 原子事件抽取结果

方法	Precision	Recall	F1-measure
None	28.81%	29.12%	28.97%
PreBi	38.46%	38.22%	38.34%
PreTri	31.42%	31.65%	31.54%
DisBi	40.97%	39.34%	40.14%
PreBi+PreTri	40.74%	40.47%	40.60%
PreBi+DisBi	41.59%	40.46%	41.02%
PreTri+DisBi	40.94%	39.82%	40.37%
All	41.64%	40.80%	41.22%

从表7可以看出,在评价原子事件整体抽取效果时,未采用任何标签搭配规则的“方法None”效果明显亦低于其他方法,从“方法PreBi”和“方法PreBi+DisBi”的对比以及“方法PreTri”和“方法PreTri+DisBi”的对比来看，新发现规则对于原子事件整体抽取效果也是有提升的.效果最好的是采用了预定义二元标签搭配规则、预定义三元标签搭配规则和新发现规则的“方法All”,F值为41.22%.

通过对抽取错误的实例进行分析,主要分为三种情况:误匹配、过匹配和欠匹配.误匹配表现为将非原子事件成分识别为原子事件成分和未能将原子事件成分识别出来两种.过匹配表现为自动识别的原子事件成分在句中的范围过度覆盖了标注语料中原子事件成分的范围,例如句子“二氧化碳是温室气体之一”,在标注语料中原子事件客体成分为“温室气体”,而本文方法自动抽取的原子事件客体成分为“温室气体之一”.欠匹配表现为自动识别的原子事件成分过短，没有恰好覆盖标注语料中的原子事件成分,例如句子“他是苏联援华抗日志愿军武汉大会战战地故迹重游团的成员”,在标注语料中原子事件客体成分为“苏联援华抗日志愿军武汉大会战战地故迹重游团的成员”,而本文方法自动抽取的原子事件客体成分为“战地故迹重游团的成员”.上述三种错误抽取的原因在于汉语是一种灵活的语言,尽管采用统计和规则相结合的方法来处理,也很难完全覆盖语用可能性,比如将动词作为名词使用、名词作为动词使用等,而且部分错误抽取的情况并不是严格意义上的错误,比如上述例句“二氧化碳是温室气体之一”,无论是“温室气体”还是“温室气体之一”作为原子事件客体成分都可以完整的描述一个原子事件.

综上所述,从表 5 至表 7 可以得出以下结论:将标签搭配规则应用于原子事件抽取是有效的,能够提升原子事件抽取的准确率、召回率和 F 值;基于 N 元模型的标签搭配规则验证与发现方法是有效的,通过加入新发现

的规则,对原子事件主体成分的抽取、原子事件客体成分抽取以及原子事件抽取都是有帮助的.

6 结论

本文以 N 元模型为基础,利用统计与规则相结合方式实现了标签搭配规则的验证与自动发现方法,完备了原子事件抽取中用于标签搭配的规则库,并通过标签搭配迭代融合方法进行原子事件抽取,实验结果表明该方法是有效的.此外,一旦得到了最终标签搭配规则库,使用标签搭配方法进行原子事件抽取不需要对文本进行句法分析、语义分析,直接在词法分析的基础之上根据标签搭配规则进行原子事件成分融合,这在一定程度上提升了原子事件抽取的效率.下一步工作计划可以尝试在原子事件抽取的基础上开展事件关系与事件语义角色标注的研究工作.

References:

[1] Liu MF, Li Y, Ji DH. Event semantic feature based Chinese textual entailment recognition. Journal of Chinese Information Processing,2013,05:129-136(in Chinese with English abstract).[doi: 10.3969/j.issn.1003-0077.2013.05.018]

[2] Zhou DY, Chen LY, He YL. A simple Bayesian modelling approach to event extraction from Twitter. In: Toutanova K,eds. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014:700-705.

[3] Glavaš G, Šnajder J. Event-centered information retrieval using kernels on event graphs. In: Kozareva Z, eds. Proceedings of TextGraphs-8 Graph-based Methods for Natural Language Processing. Seattle: Association for Computational Linguistics, 2013: 1-5.

[4] Ding X, Zhang Y, Liu T, Duan JW. Using structured events to predict stock price movement: An Empirical Investigation. In: Moschitti A, eds. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1415-1425.

[5] Lee H, Recasens M, Chang A, Surdeanu M, Jurafsky D. Joint entity and event coreference resolution across documents. In: Tsujii J, eds. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012: 489-500.

[6] Xiong NX, Vasilakos AV, Yang LT, et al. Comparative analysis of quality of service and memory usage for adaptive failure detectors in healthcare systems. In: Vasilakos AV, eds. IEEE Journal on Selected Areas in Communications. IEEE Press, 2009.27(4):495-509.

[7] Boros E, Besançon R, Ferret O, Grau B. Event role extraction using domain-relevant word representations. In: Moschitti A, eds. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1852-1857.

[8] Li Q, Ji H, Huang L. Joint event extraction via structured prediction with global features. In: Fung P,eds. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia: Association for Computational Linguistics, 2013: 73-82.

[9] Li PF, Zhou GD, Zhu QM, Hou LB. Employing compositional semantics and discourse consistency in Chinese event extraction. In: Tsujii J, eds. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012: 1006-1016.

[10] Ding X, Song F, Qin B, Liu T. Research on typical event extraction method in the field of music. Journal of Chinese Information Processing,2011,25(2):15-20(in Chinese with English abstract).[doi:10.3969/j.issn.1003-0077.2011.02.003]

[11] Xiao S, He YX. Event Hypergraph Model and Event Type Recognition. Journal of Chinese Information Processing,2013,27(1): 30-38(in Chinese with English abstract).[doi:10.3969/j.issn.1003-0077.2013.01.005]

[12] Llorens H, Saquete E, Navarro-Colorado B. TimeML events recognition and classification: learning crf models with semantic roles. In: Huang CR, eds. Proceedings of the 23rd International Conference on Computational Linguistics. Association for Computational Linguistics.Beijing: Coling 2010 Organizing Committee, 2010: 725-733.

[13] Li PF, Zhu QM, Diao HJ, Zhou GD. Joint modeling of trigger identification and event type determination in chinese event extraction. In: Kay M, eds. Proceedings of COLING 2012. Mumbai: The COLING 2012 Organizing Committee,2012:1635-1652.

[14] Li PF, Zhu QM, Zhou GD. Employing event inference to improve semi-supervised chinese event extraction. In: Junichi T, eds. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin: Dublin City University and Association for Computational Linguistics, 2014:2129-2139.

[15] Chen C, Ng V. Joint modeling for chinese event extraction with rich linguistic features. In: Kay M, eds. Proceedings of COLING 2012. Mumbai: The COLING 2012 Organizing Committee,2012:529-544.

[16] Yeh JF, Li SF, Wu MR, Chen WY, Su MC. Chinese word spelling correction based on n-gram ranked inverted index list. In: Yu LC, eds. Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing. Nagoya: Asian Federation of Natural Language Processing, 2013: 43-48.

[17] Noji H, Mochihashi D, Miyao Y. Improvements to the bayesian topic n-gram models. In: Yarowsky D, eds. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: Association for Computational Linguistics, 2013: 1180-1190.

[18] Che WX, Guo J, Liu T. Reliable dependency arc recognition. Expert Systems with Applications, 2014, 41(4): 1716-1722.

[19] Tang DY, Wei FR, Yang N, Zhou M, Liu T, Qin B. Learning sentiment-specific word embedding for twitter sentiment classification. In: Toutanova K,eds. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics, 2014: 1555-1565.

[20] Shan YX, Chen X, Shi YZ, Liu J. Fast language model look-ahead algorithm using extended n-gram model. Acta Automatica Sinica, 2012, 38(10): 1618−1626(in Chinese with English abstract).[doi:10.3724/SP.J.1004.2012.01618]

[21] Gong ZX, Zhang M, Tan C, Zhou GD. N-gram-based tense models for statistical machine translation. In: Tsujii J, eds. Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics, 2012: 276-285.

[22] Xia F. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). http://repository.upenn.edu/ircs_reports/38/. 2000.

附中文参考文献:

[1] 刘茂福,李妍,姬东鸿. 基于事件语义特征的中文文本蕴含识别[J]. 中文信息学报,2013,05:129-136.

[10] 丁效,宋凡,秦兵,刘挺.音乐领域典型事件抽取方法研究[J].中文信息学报,2011,25(2):15-20.

[11] 肖升,何炎祥.事件超图模型及类型识别[J].中文信息学报,2013,27(1):30-38.

[20] 单煜翔,陈谐,史永哲,刘加.基于扩展 N 元文法模型的快速语言模型预测算法[J].自动化学报,2012,38(10):1618-1626.