

# 基于新闻事件片段的时序关系识别方法

李英俊, 张宏莉, 王星

(哈尔滨工业大学计算机科学与技术学院, 黑龙江 哈尔滨 150001)

**摘要:** 针对新闻文本, 将事件片段定义为最小划分单元, 提出了一个两阶段的事件片段时序关系识别算法, 第一阶段时序映射和第二阶段时序识别。使用真实网上新闻数据进行实验, 结果表明, 所提方法相比普通的分类器及强规则算法, 准确率提高了 4%。

**关键词:** 事件片段; 时序映射; 时序关系识别; 全局优化

**中图分类号:** TP393

**文献标识码:** A

**doi:** 10.11959/j.issn.2096-109x.2017.00171

## Temporal relation recognition method based on news event fragments

LI Ying-jun, ZHANG Hong-li, WANG Xing

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

**Abstract:** Event fragments as the smallest unit was defined and a two-stage event fragments temporal relation recognition algorithm was proposed. The first stage is sequential mapping and the second is sequence recognition. The real online data increased accuracy by 4% compared to common classifier and strong rule algorithm.

**Key words:** event fragments, sequential mapping, temporal relation recognition, global optimization

### 1 引言

随着搜索引擎分析技术的提高、数据挖掘技术的发展, 事件时序识别逐渐成为研究热点。事件是指在特定时间和环境下, 某种行为或状态的描述<sup>[1]</sup>。从时间角度上看, 事件可能发生在一段时间区间或持续一段时间。因此, 事件之间会存在一定的先后顺序。事件的时序关系识别是根据事件时间线索特征进行事件时间位序识别。对于新闻类的文本而言, 其中包含多个事件和时间, 通过事件-时间的对应以及事件时序识别, 可以更

好地分析理解文本内容。

例如, 一篇新闻一般除了包含一个主题外, 还包含多个事件, 如表 1 中的例 1 所示, 文本是节选自一篇新闻, 事件分别用“[ ]”标示出来, 对应的发生时间用下划线标出。从一篇新闻里抽取多个事件, 并对这些事件发生时间进行排序, 即事件时序关系识别。现有研究主要集中在句内事件-时间关系和相邻句中主要事件的时间关系。句内事件-时间关系识别是对同一个句子中事件和时间表达式关系识别, 如表 1 中的例 2 所示, “投资”和“近年”的关系是有交叉或投资发

**收稿日期:** 2017-04-07; **修回日期:** 2017-05-10. **通信作者:** 李英俊, lyjamare@163.com

**基金项目:** 国家自然科学基金资助项目 (No.61402137); 国家重点基础研究发展计划 (“973”计划) 基金资助项目 (No. 2013CB329602)

**Foundation Items:** The National Natural Science Foundation of China (No. 61402137), The National Basic Research Program of China (973 Program) (No. 2013CB329602)

生于近年之后；相邻句中主要事件的时间关系识别是指对相邻句子中 2 个事件词之间的时序关系识别，一般默认这个事件词是句子中的主要动词，如表 1 中的例 3，“撤离”和“结束”的关系是撤离先于结束发生。

表 1 事件时序识别例子

序号	例句
例 1	……「今天，由中国贸易促进会广东分会与香港中华总商会联合主办的“九六税改及进口原料台账制执行实务研讨会”在广州举行，向客商介绍「今年国家将实行的外经贸三大政策调整。」」……
例 2	近年来，江苏省在对外开放中下大力气进一步改善投资环境，获得重大进展，增强了对外商投资的吸引力
例 3	美、英、法三国的士兵也将于 9 月初最后撤离柏林。德国领土上存在占领军的状态行将结束

针对上述问题，使用事件片段作为最小分析单元来表征一个事件。事件片段是将一篇新闻划分为时间上独立的文本片段，这段文本片段由一句或几句话组成，并且叙述的事情都是发生在一个时间区间，如例 1，包含 2 个事件片段，一个是以“今天”作为时间区间，事件片段是一整句话，另一个是以“今年”作为时间区间，事件片段是“今年……调整”这句话。这种划分既保证事件的完整性也保证其独立性。本文首先给出事件片段和事件时序识别的定义，提出了一个两阶段算法 two-stage-LTP 算法（以下简称 tsLTP 算法），第一步采用有监督的机器学习算法，将一篇文章划分为多个事件片段，然后再将识别出的时间区间与事件片段对应起来。第二步采用整数线性规划的全局优化推理模型，得到事件片段间的时序关系。最后，构建实际数据集，并将本文方法和现有的基准方法进行对比，实验表明，本文方法在准确率、召回率、平均准确率和 F1 评分方面都有较大的提高。文本的研究成果不仅可以帮助用户获取到事件脉络，也可以应用在舆情领域、趋势演化中。

## 2 相关工作

事件时序识别的方法分为局部时序关系识别和全局时序关系识别两类，局部时序关系识别只是识别同一个句子内的 2 个事件词或事件-时间关系，和句内事件-时间关系识别任务相似；全局

时序关系识别除同一个句子内的事件关系识别，也对相邻句子以及跨句子之间的 2 个事件词关系进行识别。

### 2.1 局部时序关系识别

局部时序关系识别早期的研究是基于特殊规则的方法，由于规则的方法没有普适性，而后学者们提出了使用机器学习的方法进行时序关系识别。

研究学者使用规则的方法，一般是基于语言学知识、上下文特征定义好规则集，再进行时序关系的分析和识别。Allen<sup>[3,4]</sup>将时序关系看作是事件时间信息之间的比较，提出了 13 种时序关系，利用比较后生成的时间区间关系定义了时序关系的推理规则。文献[5~9]利用对时间关系影响的语言学知识和上下文特征，如时态、修辞关系、语用约束、实际惯例和背景知识建立规则。Chklovski 等<sup>[10]</sup>定义了 6 种时序关系，利用触发词建立词-句匹配模板抽取事件时序关系对，由于规则的方法具有主观性以及强制适配性。虽然规则的方法识别结果准确率高，但召回率不够理想，没有实际应用价值。

研究学者使用机器学习的方法是把时序关系识别看作分类问题，将文本片段中抽取出的实体-时间对表示为特征向量，选择适当的特征和机器学习分类模型进行分类。TimeML<sup>[11]</sup>是国际标准化组织（ISO, International Organization for Standardization）为标注事件和时间表达式而定义的一套标注体系，TimeBank<sup>[12]</sup>语料库是根据 TimeML 标注体系进行标注的语料库。大多数研究学者利用这个语料库研究事件时序关系识别。使用的特征可以分为词性特征，如时间词、介词、连词等与表征时间关系的词；事件特征，包括时态、体态和情态；句法特征，如事件和时间构成的句法分析树；语义特征，如句子中事件动词的语义角色、事件共指关系。Mani 等<sup>[13]</sup>在假设 TimeBank 语料库抽取的词性语义信息全部正确的基础上，使用最大熵分类模型对事件对——时序关系识别。Chambers 等<sup>[14]</sup>在 Mani 的基础上，加入了更多的事件、句法特征，并且提出了一个两阶段的时序关系识别方法，第一阶段是提取事件属性，第二阶段是把第一阶段的结果作为

特征来识别事件之间的时序关系,实验效果比Mani好一些。Bethard等<sup>[15]</sup>提出了一种结合句法关系和语义特征识别时序关系的方法。Souza等<sup>[16]</sup>提出了篇章级的特征,对系统性能的提高有一定的帮助。Mirza等<sup>[17]</sup>通过实验发现使用一部分简单的特征就可以获取较好的效果。Li等<sup>[18]</sup>参考了部分英文事件时序识别常用的特征,使用贝叶斯分类的机器学习方法对中文事件时序关系识别。Cheng等<sup>[19]</sup>借鉴TimeML标注准则,构建了一个中文事件时序的语料库,同时利用事件的类型来识别中文事件时序关系。王风娥<sup>[20]</sup>采用最大熵分类器识别句子内的事件时序关系,实验语料利用了TempEval-2提供的中文语料库。以上研究表明,不同的特征、不同的时序识别目的以及不同的机器学习方法都会对实验结果产生不同的影响。

局部时序关系识别更多地使用了词法、句法语义特征,可以很好地判断2个词语的时序关系,然而这些特征却不能直接用于判断两句话的时序关系。本文提出的方法则在事件-时间对应的基础上,加以全局优化的方法,可以识别出句子级别的事件片段间的时序关系。

## 2.2 全局时序关系识别

机器学习方法解决了事件对之间的时序关系识别问题,而且效率也很高。但时序识别的最终目的是全局的时间关系,也就是时间链。因此,研究人员在机器学习方法的基础上,提出了全局推理模型来识别。

Chambers等<sup>[21]</sup>在分类器结果的基础上,使用整数线性规划方法提高实验性能。Denis<sup>[22]</sup>通过用区间端点表示事件和时间的方法控制推理方法的复杂性,并提出一种图分解方法简化

图的优化问题。Do<sup>[23]</sup>提出一种联合推理的事件时间链结构,即把一篇文章中的事件按照其发生的时间先后顺序构造成一条完整的事件链,然后,使用整数线性规划模型优化该事件链。另外,Do<sup>[23]</sup>还加入事件同指关系,使实验结果有了显著的提高。Ng<sup>[24]</sup>在Do<sup>[23]</sup>的基础上借助篇章结构关系方面的技术进一步提升了实验的性能。而Yoshikawa<sup>[25,26]</sup>则使用马尔可夫网络推理模型提高机器学习方法的性能。然而,其中大多数使用的全局优化模型的约束条件较单一,仅包含反转性、传递性、时间表达式信息等规则。Chang等<sup>[27]</sup>提出了一个生活圈模型,对用户的推特数据进行分析,生成时间链。

全局模型解决了跨句子之间2个事件词的时序关系,然而现有全局模型中的约束条件单一,还是主要依赖机器学习的结果。本文方法提出的约束条件,可以优化句子间的关系,提高全局识别结果。

## 3 问题定义与描述

本文研究的问题是将一篇新闻文章划分为多个事件片段,得到事件片段之间的时序关系,如图1所示,将一篇新闻划分为多个事件片段。 $\{e_1, e_2, \dots\}$ ,然后将文章中抽取的时间和事件片段对应,最后进行事件之间的时序识别,得到事件之间的时序关系。

**定义1** 时间区间 $t$ ,包含起始时间 $t_B$ 和结束时间 $t_E$ ,  $t=[t_B, t_E]$ 。

**定义2** 事件片段为 $e$ ,叙述的事情发生在一个时间区间内。一篇文章相当于事件片段集合 $E$ ,每个 $e \in E$ ,  $e=\{s_1, s_2, \dots, s_n, t\}$  ( $n \geq 1$ ),其中, $s_i$  ( $i \in [1, n]$ )表示构成事件片段 $e$ 的一个子句。

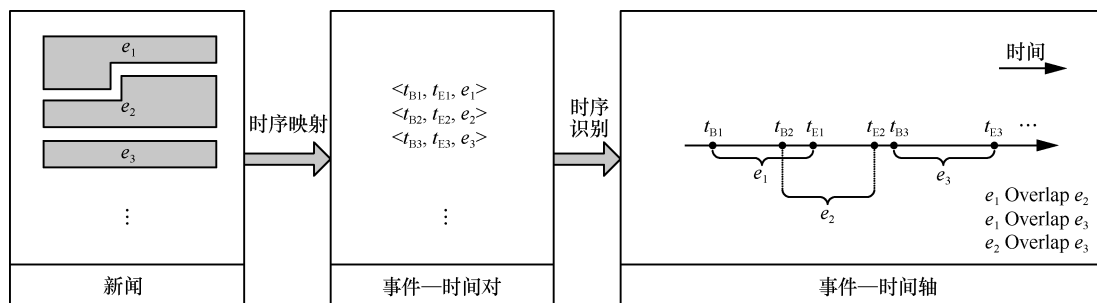


图1 问题定义

**定义3** 时序关系为 Before、After、Overlap、Include、Simultaneous、None，具体描述如表2所示。

表2 时间关系定义

关系	描述
Before	$ts_i$ 先于 $ts_j$ 发生
After	$ts_i$ 后于 $ts_j$ 发生
Overlap	$ts_i$ 和 $ts_j$ 的发生持续时间有交叉
Include	$ts_i$ 的发生持续时间区间包含 $ts_j$
Simultaneous	$ts_i$ 和 $ts_j$ 同时发生
None	$ts_i$ 和 $ts_j$ 时序关系未知

由上述定义可知，表1的例1中有2个事件  $e_1$  和  $e_2$ ，分别对应2个事件区间  $t_1$  和  $t_2$ ，其中  $e_1$  和  $e_2$  的时序关系是同时发生，具体描述如下。

$t_1 = \text{今天} \{2013 \text{ 年 } 4 \text{ 月 } 10 \text{ 日}, 2013 \text{ 年 } 4 \text{ 月 } 10 \text{ 日}\}$

$t_2 = \text{今年} \{2013 \text{ 年 } 1 \text{ 月 } 1 \text{ 日}, 2013 \text{ 年 } 12 \text{ 月 } 31 \text{ 日}\}$

$e_1 = \{\text{由中国贸易促进会广东分会与香港中华总商会联合主办的“九六税改及进口原料台账制执行实务研讨会在广州举行，向客商介绍今年国家将实行的外经贸三大政策调整。今天}\}$

$e_2 = \{\text{国家将实行的外经贸三大政策调整。今年}\}$

$e_1 \text{ Overlap } e_2$

## 4 tsLTP 算法

本文提出一个两阶段算法，第一阶段是时序映射，第二阶段是时序识别。时序映射包含时间提取、事件划分和事件-时间对应；时序识别包含机器学习识别获得初始识别结果，全局优化模型得到最优识别结果。算法流程如图2所示。

### 4.1 时序映射

#### 1) 时间提取

使用复旦大学的时间识别库进行时间提取。

#### 2) 事件划分

一个事件片段是由多个句子构成的，汉语句子中最小独立单元是以逗号间隔的句子，因此，在识别完新闻中包含的时间信息后，对每个句子进行分析、构造事件片段，继而完成事件片段的划分，利用机器学习的方法，选取特征如表3所示。判断2个最小单元是否属于同一事件片段。

#### 3) 事件-时间对应

一个事件片段往往对应一个时间，本文使用如下的规则确定时间与事件片段的对应关系。

① 根据语言学规律，时间词语的管辖方式一般都是后向管辖，因此，默认一个事件片段的的时间区间是与其最近的时间区间。

② 如果事件片段最近的是另外一个事件片段，则选择文档创建时间作为此事件片段的的时间区间。

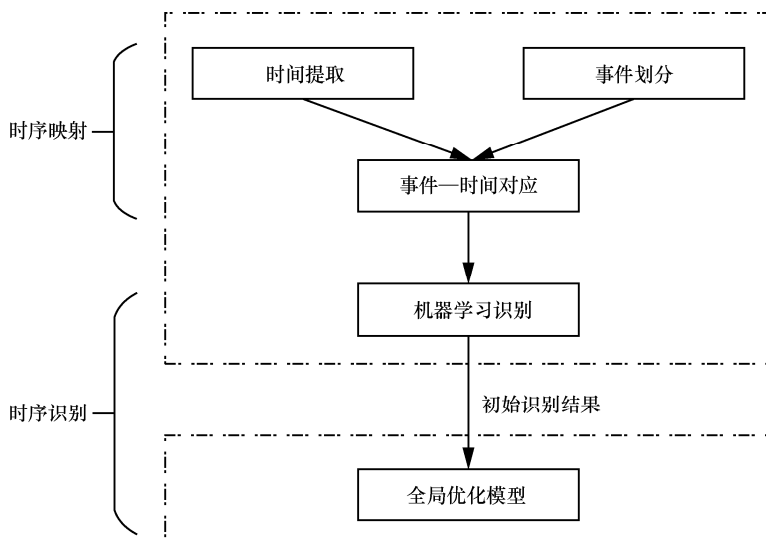


图2 算法流程

表3 事件片段划分特征集合

特征编号	描述
①	2个句子之间是否存在时间词, 存在: $\text{exist\_t} = 1$ ; 不存在: $\text{exist\_t} = 0$
②	2个句子之间是否存在转折词, 存在: $\text{exist\_tr} = 1$ ; 不存在: $\text{exist\_tr} = 0$
③	2个句子中间间隔句子数, 句子数: $\text{num}$
④	2个句子是否有时间功能词连接, 存在: $\text{exist\_f} = 1$ ; 不存在: $\text{exist\_f} = 0$
⑤	2个句子先后顺序, 先于: $\text{order} = \text{b}$ 后于: $\text{order} = \text{a}$

## 4.2 时序识别-机器学习

时序识别的第一步是用机器学习, 获得识别的初始识别结果, 本文也将机器学习方法作为基准系统。由于本文研究的是同一文档内任意2个事件片段之间的时序关系, 而现有方法所用的一些词法、句法等特征只限于单个句子内, 对于句子级或相邻较远的2个事件片段, 此类特征可能不会发挥太大作用。因此, 本文根据中文语言特点提出了篇章级别的特征, 大致可归纳为语义特征、特殊词、事件关系以及事件所处的上下文信息等几类特征, 如表4所示。

表4 基准系统特征集合

特征类别	具体特征
语义特征	2个事件片段核心词的语义相似度, 用 $\text{sim}$ 表示 2个事件片段核心词是否为同义, 用 $\text{if\_syn}$ 表示
特殊词	时间区间是否是 DCT, 用 $\text{if\_dct}$ 表示 时间区间先后顺序, 用 $\text{order\_t}$ 表示 时间区间时态信息先后, 用 $\text{order\_tense}$ 2个事件片段之间是否存在连接词, 用 $\text{if\_conj}$ 表示
上下文信息	2个事件片段在文中出现的先后顺序, 用 $\text{order\_a}$ 表示 2个事件片段是否在同一个句子中, 用 $\text{if\_in}$ 表示 2个事件片段间隔的句子数目, 用 $\text{num}$ 表示

针语义特征, 2个句子的核心词如果有一定的相似关系, 就可以知道这2个句子之间的时序关系。2个词相似度越高, 说明2个词叙述的事件应该是同一个, 这样即使2个句子间隔较远, 可知2个句子可能是同时或时间区间有交叉。

特殊词, 句子级别之间的时序比较, 最重要的依赖应该是特殊词, 因为特殊词基本可以得出2个句子之间的关系。2个事件有明确的时间区间, 基本就可以知道2个事件的时序关系。连接

词也可以辅助判断。

上下文关系, 中文的叙述, 很注重上下文的衔接, 因此, 这个特征也可以有助于判断事件之间的关系。

## 4.3 时序识别-全局优化模型

由于事件时序分类器只是对每一对事件片段的关系进行两两分类, 没有考虑事件片段对之间的联系, 因而会将一些本可以区分出时序关系的事件对区分错误。已有的全局优化方法仅使用自反性和传递性约束条件, 因为这2个条件是在已有识别结果上进行优化, 如果结果准确性及粒度不够, 就不能够提高优化的效果。因此, 本文在已有全局优化方法的基础上, 提出了几个约束条件。事件时序分类器具有概率特性, 而同时本文后续提出的约束条件也具有概率特性。因此, 同样可以将本文研究的内容, 转化为一个全局最优的问题, 本文使用整数线性规划方法解决这个优化问题。

### 4.3.1 目标函数

目标函数定义如式(1)所示。

$$\arg \max_x \sum_{e_i e_j \in EE} \sum_{r \in R} a \log p(e_i e_j, r) x(e_i e_j, r) \quad (1)$$

其中,  $E$  表示事件片段集合,  $R$  表示关系集合,  $x(e_i e_j, r)$  表示事件  $i$  和  $j$  的关系, 取值为1或0;  $a$  表示后续提出约束条件的概率值;  $p(e_i e_j, r)$  为事件  $i$  和  $j$  关系分类预测器的概率值。

### 4.3.2 基本约束条件

#### 1) 唯一性

唯一性确保2个事件之间只能存在一种关系, 即针对本文提出的6种关系, 2个事件之间的关系只能是其中一种, 因此,  $x(e_i e_j, r)$  只有在对应的关系上取值为1, 其他关系时, 取值为0。

$$x(e_i e_j, r) \in \{0, 1\}, \forall r \in R \quad (2)$$

$$\sum_{r \in R} x(e_i e_j, r) = 1, \forall e_i e_j \in EE \quad (3)$$

#### 2) 自反性

自反性是指当  $x(e_i e_j, r)$  取1时, 就可以知道  $x(e_i e_j, r)$  取1时对应的关系。例如, 如果  $x(e_i e_j, \text{before}) = 1$ , 则  $x(e_i e_j, \text{after}) = 0$ 。定义的6种时序关系都符合自反性。Include 的自反为

Included (被包含), 本文未提及。

$$x(e_i e_j, r) - x(e_i e_j, \hat{r}) = 0, \forall e_i e_j \in EE, i \neq j \quad (4)$$

### 3) 传递性

传递性是指时序关系的传递闭包, 即  $e_i$  和  $e_j$  具有关系  $r_a$ ,  $e_j$  和  $e_k$  具有关系  $r_a$ , 则  $e_i$  和  $e_k$  也具有关系  $r_a$ 。例如,  $e_i$  和  $e_j$  具有关系 Before, 即  $e_i$  发生于  $e_j$  之前,  $e_j$  和  $e_k$  具有关系 Before, 即  $e_j$  发生于  $e_k$  之前, 则可以利用时序关系传递闭包的性质推导出  $e_i$  和  $e_k$  也具有关系 Before, 即  $e_i$  发生于  $e_k$  之前。

$$x(e_i e_j, r) + x(e_j e_k, r) - x(e_i e_k, r) \leq 1, \\ \forall e_i e_j, e_j e_k, e_i e_k \in EE, i \neq j, j \neq k, i \neq k \quad (5)$$

#### 4.3.3 新增约束条件

基本约束条件在分类结果准确性保障的前提下才是有效的。而分类器的结果是存在概率的, 因此对 17 000 篇搜狗爬取的新闻语料统计分析后, 新增加 3 个约束条件, 分别是时间区间细粒度、时间-事件映射修正、篇章连接词。

##### 1) 时间区间细粒度

时间区间粒度越细, 2 个时间区间比较出来的结果可信度就越高。如果 2 个时间区间的粒度过大, 就降低了 2 个时间区间的准确度, 增加了时序关系的模糊度。因此可以将时间区间划分为: 年  $T_{\text{year}}$ , 月  $T_{\text{month}}$ , 日  $T_{\text{day}}$ , 上、中、下午  $T_{\text{hour}}$  这几个时间粒度不同的类型, 同时对每一个类型设定一个阈值, 称作时间区间的置信度。时间区间粒度越细, 则置信度越高。有些事件片段的确发生在某一段时间内, 虽然其对应的时间区间是一个时间段, 但可能存在时间段范围过大的问题。例如, 2012 年五一黄金周黄果树旅游景区游客暴增。事件片段对应的是 2012 年, 但可以知道这个事件指的是五一黄金周期间。通过对 17 000 篇新闻语料统计, 提取了出现次数较多的类似五一黄金周这类可以指向某一段时间的词语, 如表 5 所示。

表 5 时间区间特殊词

特殊词	时间区间
春节	过年的 7 天
五四	5 月 4 日
六一	6 月 1 日
圣诞节	12 月 24 日至 12 月 26 日
国庆节	10 月 1 日至 10 月 7 日

所以, 除了对时间区间的粒度进行阈值修正处理, 还要对这一类有明确指向性的词语进行时间区间修正。

$$\text{iff } T_{t_i} = T_{t_j} \quad a = 1, \text{ else } a = 0.5 \quad (6)$$

其中,  $T_{t_i}, T_{t_j}$  表示  $e_i, e_j$  的时间区间细粒度。

##### 2) 时间-事件映射修正

通过已有统计, 只有 60% 的文章存在时间区间, 而且并不是每篇文章的每个句子间都存在时间区间。基于新闻语料的统计, 发现有一类显示的时序功能词, 分别为时间副词、关系连词以及功能词组合, 这类功能词, 可以表示共时、顺序和逆序关系。

共时关系是指事件发生时间具有共时特性。例如, “同时” 和 “并” 这 2 个词都可以明显地看出来, 表明词语前后事件指明的发生时间是共时的。共时关系的词一般位于 2 个事件中间。

顺序关系是指事件发生时间具有顺序特性, 即事件发生具有先后次序。例如, “首先…其次…” 这对连词, 连接的 2 个事件具有顺序关系。

逆序关系是指事件发生时间具有逆序特性, 即文字中事件的叙述先后次序和真实发生的次序刚好相反。

因此, 一般新闻类事件, 开头多为时间区间, 而后会使用这类时序功能词完成一段话或一个事件的描述。根据本文事件片段划分的方式, 连接词会将事件划分为 2 个事件片段。此时, 利用时序功能词推出事件片段与前后事件的关系, 继而利用基本约束条件中的传递性, 可以识别出全局事件时序关系。

$$\text{iff } \exists \text{conj\_t between } e_i \text{ and } e_j \text{ then } x(e_i e_j, r) = 1 \quad (7)$$

其中, conj\_t 表示时序功能词,  $r$  表示 conj\_t 对应的时序关系。

##### 3) 篇章连接词

在汉语篇章中, 句子与句子或子句与子句之间存在语义连接关系, 如条件关系、转折关系、因果关系等, 连接词主要指连接不同单位并表示这种语义关系的词语。根据哈尔滨工业大学中文篇章关系语料库的统计, 连接词的关系大致可以分为因果类、并列类、转折类、解说类。将这 4 类篇章连接词对应到时序关系上, 如因果连接词, 表明 2 个事件是有先后顺序的。

尽管篇章连接词可以表征时序关系，然而有些连接词却可以表示多种关系，如“而”这个连接词，既可以表示递进关系、顺承关系，也可以表示转折关系和并列关系。这些关系对应的时序关系是不同的，可以是先后顺序、同时发生和没有关系这3类。因此，根据统计结果，对每个可以表示多个关系的连接词设定概率值，通过每个连接词的统计结果表示概率值。

iff  $\exists \text{conj\_p}$  between  $e_i$  and  $e_j$  then  $x(e_i e_j, r) = a$  (8)

$a = r\_conj\_p / \text{relation\_conj\_p}$  (9)

其中， $\text{conj\_p}$  表示篇章连接词， $r$  表示  $\text{conj\_p}$  对应的时序关系，连接词表示  $r$  关系这一事件的概率用  $a$  表示。

## 5 实验

### 5.1 实验数据集

本文实验所用的语料是从网络上主流新闻媒体爬取下来的，包含144篇新闻报道，预处理使用上文提到的方法进行了事件片段划分、时间-事件对应。同时对任意2个事件片段的时序关系进行人工标注。实验语料由2位标注者共同完成，标注结果Kappa值达到0.85。Kappa值常用于协同标注一致性判断，大于0.8即可认为完全一致，达到0.85说明2位标注者对标注规范理解是一致的。共计抽取出898个时间区间和划分出1113个事件片段以及标注16084个时序关系。实验采用十折交叉的方法，定义评价标准指标准确率  $P(\text{precision})$ 、召回率  $R(\text{recall})$  及  $F1$  值。通过比较算法得到分类结果  $TP$  和标注的结果  $TN$ ， $P$ 、 $R$ 、 $F1$  计算如下。

$$P = \frac{TP \cap TN}{TP}, R = \frac{TP \cap TN}{TN} \quad (10)$$

$$F1 = \frac{2PR}{P + R} \quad (11)$$

现有时序识别方法基本采用相同的机器学习算法流程，即将每一事件对构造成向量形式，使用监督学习的方式进行时序识别，本文实验采用4.2节中提到的特征，用于监督学习算法成为一个基准系统，以此作为方法一。根据引用文献[2]中提出的规则，实现了一个强规则推理的方法，作为方法二。与本文方法较为相似的Do等提出的JIETC方法作为方法三，本文也做了实验进行

对比。

### 5.2 实验结果与分析

实验结果如表6所示，本文研究的是事件片段之间的时序关系，因此，对比Do的方法主要是其提出的全局优化方法。

表6 各方法实验性能比较

方法	$P$	$R$	$F1$
方法一	72.2%	72.34%	72.27%
方法二	73.31%	73.41%	73.36%
方法三	74.19%	74.32%	74.25%
本文方法	79.76%	80.01%	79.88%

首先，比较准确率这个指标，本文方法比其他最高值高出5%，由于本文的方法是在机器学习的基础上做了全局优化，不仅考虑自反性、传递性这些基础约束条件，而且针对新闻事件描述特性，加入了3个约束条件，有效地对结果进行了优化，提高了识别的准确率。

其次，比较召回率这个参数，由于强规则推理模式，依赖数据程度较高，因此召回率会比其他算法低10%左右；由基准系统和Do方法可知，虽然Do方法是在机器学习的结果上使用传递性和自反性进行约束修正结果，但召回率只相差2%，说明单独机器学习中的结果存在差错，导致自反性和传递性的优化存在错误；本文提出的3个新增约束条件，考虑了一些引起错误的条件，进行约束优化，如事件-时间修正这一条件，可以把时间这一项修正为正确项，自然会提高识别结果。

再对每个约束条件进行性能对比，如图3所示，可以发现，传递性对全局时序关系的识别，有一定影响和提高。比较本文提出的3个约束条件，可以看出时间-事件映射关系的修正以及篇章连接词有效地提高了时序识别的准确率，本文认为，错误的时间与事件对应的确导致了机器学习算法的识别准确率，通过修正，提高机器学习的准确率，加之传递性，使全局识别率提高很多。同理，篇章连接词的提出可以保证相邻2个句子的时序关系准确性，继而利用传递性，可以更准确地识别2个事件的时序关系。

本文实验所用新闻文本涉及领域广泛，并没有局限在某一领域，如医药、金融等。所以具有较高的扩展性和通用性。

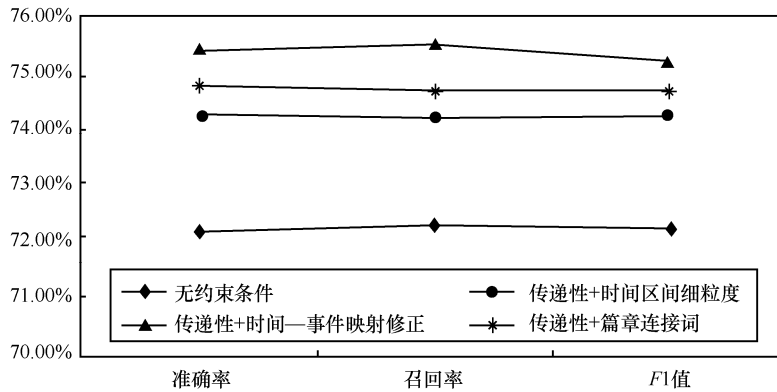


图3 本文提出的约束条件性能对比

## 6 结束语

本文提出了一种新的事件时序识别方法，以事件片段作为最小粒度研究单元，保证了事件的完整性，避免了只以动词或名词构成事件作为研究单元产生的不合理性和时间区间稀疏的问题。在片段划分、时间区间与事件片段一一对应的基础上，在事件片段时序识别方面，优化全局识别公式并提出多个有效的约束条件，提高了时序识别的准确率。

在下一步工作中，将文字类叙述时间的知识库建立起来，使这一部分的时间区间可以识别出来；另外，引入叙述类文本模板，将叙述类文本中的事件片段识别划分的准确率提高。

## 参考文献：

- [1] 郑新, 李培峰, 朱巧明, 等. 中文事件时序关系的标注和分类方法[J]. 计算机科学, 2015, 42(7): 276-279.  
ZHENG X, LI P F, ZHU Q M, et al. Annotation and classification of temporal relation between chinese events[J]. Computer Science, 2015, 42(7): 276-279.
- [2] 庞黎多. 基于时间片段的时间关系识别系统的设计与实现[D]. 太原: 山西大学, 2014.  
PANG L D. System of temporal relation based on temporal segmentation[D]. Taiyuan: Shanxi University, 2014.
- [3] ALLEN J F. Maintaining knowledge about temporal intervals[J]. Communications of the ACM, 1983, 26(11): 832-843.
- [4] ALLEN J F. Towards a general theory of action and time[J]. Artificial Intelligence, 1984, 23(2): 123-154.
- [5] PASSONNEAU R J. A computational model of the semantics of tense and aspect[J]. Computational Linguistics, 1988, 14(2): 44-60.
- [6] WEBBER B L. Tense as discourse anaphor[J]. Computational Linguistics, 1988, 14(2): 61-73.
- [7] LASCARIDES A, ASHER N. Temporal interpretation, discourse relations and commonsense entailment[J]. Linguistics and Philosophy, 1993, 16(5): 437-93.
- [8] HITZEMAN J, MOENS M, GROVER C, et al. Algorithms for analysing the temporal structure of discourse[C]//The Seventh Conference on European Chapter of the Association for Computational Linguistics. 1995.
- [9] KEHLER A. Resolving temporal relations using tense meaning and discourse interpretation[C]//Formalizing the Dynamics of Information. 2000.
- [10] CHKLOVSKI T, PANTEL P. Global path-based refinement of noisy graphs applied to verb semantics[C]//The International Conference on Natural Language Processing. 2005: 792-803.
- [11] SAURÍ R, LITTMAN J, GAIZAUSKAS R, et al. TimeML annotation guidelines, version 1.2.1[C]//Event London. 2006.
- [12] PUSTEJOVSKY J, HANKS P, SAURÍ R, et al. The timebank corpus[C]//Conference on Corpus linguistics. 2003.
- [13] MANI I, VERHAGEN M, WELLNER B, et al. Machine learning of temporal relations[C]//The 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. 2006: 753-760.
- [14] CHAMBERS N, WANG S, JURAFSKY D. Classifying temporal relations between events[C]//The 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. 2007: 173-176.
- [15] BETHARD S, MARTIN J H. CU-TMP: Temporal relation classification using syntactic and semantic features[C]//The 4th International Workshop on Semantic Evaluations. 2007: 129-132.
- [16] D'SOUZA J, NG V. Classifying temporal relations with rich linguistic knowledge[C]//HLT-NAACL. 2013.
- [17] MIRZA P, TONELLI S. Classifying temporal relations with simple features[C]//EACL. 2014.
- [18] LI W, WONG K F, CAO G, et al. Applying machine learning to Chinese temporal relation resolution[C]//The 42nd Annual Meeting on Association For Computational Linguistics. 2004: 582-588.



- [19] CHENG Y, ASAHARA M, MATSUMOTO Y. Constructing a temporal relation tagged corpus of Chinese based on dependency structure analysis[C]//The 14th International Symposium on Temporal Representation and Reasoning. 2007:59-69.
- [20] 王风娥, 谭红叶, 钱揖丽. 基于最大熵的句内时间关系识别[J]. 计算机工程. 2012, 38(04): 37-9.  
WANG F E, TAN H Y, QIAN Y L. Recognition of temporal relation in one sentence based on maximum entropy[J]. Computer Engineering, 2012, 38(4): 37-9.
- [21] CHAMBERS N, JURAFSKY D. Jointly combining implicit constraints improves temporal ordering[C]// The Conference on Empirical Methods in Natural Language Processing. 2008: 698-706.
- [22] DENIS P, MULLER P. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition[C]// IJCAI-11-International Joint Conference on Artificial Intelligence. 2011: 1788-1793.
- [23] DO Q X, LU W, ROTH D. Joint inference for event timeline construction[C] //The Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. 2012: 677-687.
- [24] NG J P, KAN M Y, LIN Z, et al. Exploiting discourse analysis for article-wide temporal classification[J]. Annt Journal, 2013, 11(2): 9-26.
- [25] YOSHIKAWA K, RIEDEL S, ASAHARA M. Jointly identifying temporal relations with markov logic. [C]//The Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language. 2009: 9-16.
- [26] YOSHIKAWA K, ASAHARA M, IIDA R, et al. Identifying temporal relations by sentence and document optimizations joint inference for event timeline construction[C]//Association for Computational Linguistics. 2012: 677-687.
- [27] CHANG Y, TANG J, YIN D, et al. Timeline Summarization from social media with life cycle models[C]// The Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16). 2016.

#### 作者简介:



**李英俊** (1991-), 男, 黑龙江大庆人, 哈尔滨工业大学硕士生, 主要研究方向为网络舆情、网络安全。



**张宏莉** (1973-), 女, 吉林榆树人, 博士, 哈尔滨工业大学教授、博士生导师, 主要研究方向为网络与信息安全、网络测量与建模、网络计算、并行处理等。



**王星** (1981-), 男, 黑龙江哈尔滨人, 博士, 哈尔滨工业大学助理研究员, 主要研究方向为信息安全、网络舆情、知识迁移等。