文章编号：1003-0077（2011）00-0000-00

使用全局优化方法识别中文事件因果关系*

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：分类器模型是目前识别因果关系的主要模型，该方法存在的问题是只考虑两个事件之间的关系，没有考虑同一文档中其它关联事件所包含的信息，识别结果往往存在逻辑矛盾。本文提出了一个中文事件因果关系识别的全局优化方法，该方法采用整数线性规划（ILP）的推理方法，对基本逻辑关系、因果标志词、事件类型、论元信息进行有效约束，以文档为单位来优化因果关系的识别。在本文标注的语料上的实验结果表明，与分类器方法相比，本文提出的全局优化方法的 F1 值提升了 5.54%。

关键字：事件关系；因果关系；整数线性规划；全局优化

中图分类号：TP391　　**文献标识码：**A

Using Global Optimization Method to Recognize Causal Relation between Events

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: Commonly, classifier-based model is widely applied to identify causal relation between events. One issue of this model is that it only considers the relationship between two event mentions and ignores their relevant event mentions in the same document. Thus, the results may exist many logical contradictions. This paper proposes a global optimization approach to recognize causal relation between events, which used an inference method based on Integer Linear Programming (ILP). This approach introduces various kinds constraints, i.e. the basic logical relationship, the causal signal words, the events type, the arguments information, to improve the performance. The experimental results on our annotated corpus show that our global optimization approach improves the F1-scoure by 5.54%, compared with a classifier-based model.

Key words: Event Relation; Causal Relation; Integer Linear Programming; Global Optimization

1 引言

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。一篇文章的通常由多个事件组成，这些事件表达了文章的核心内容。这些事件由某一主题将其串联起来，事件之间绝不是孤立存在，而是相互联系存在的。

识别事件关系是篇章理解的重要内容。事件之间存在多种关系，如因果关系、时序关系等。其中，因果关系是较为复杂的关系，它不仅是语言学概念，也是理解知识后进行推理的

* **收稿日期：**　　**定稿日期：**

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

作者简介：黄一龙（1991——），男，硕士研究生，中文信息处理；李培峰（1971——），男，教授，中文信息处理；朱巧明（1963——），男，教授，中文信息处理。

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

过程。识别事件因果关系，有利于获取文章主旨，有助于篇章理解、文本摘要、自动问答任务，也可用于对事件未来发展趋势做预测。

分类器模型是目前识别因果关系的主要模型，该方法存在的问题是只考虑两个事件之间的关系，没有考虑同一文档中其它关联事件所包含的信息，识别结果往往存在逻辑矛盾。本文提出了一个中文事件因果关系识别的全局优化方法，该方法采用整数线性规划（Integer Linear Programming, ILP）的推理方法，对基本逻辑关系、因果标志词、事件类型、论元信息进行有效约束，以文档为单位来优化因果关系的识别。在本文标注的语料上的实验结果表明，与分类器方法相比，本文提出的全局优化方法的 F1 值提升了 5.54%。

本文结构如下：第 2 章介绍相关工作；第 3 章介绍中文事件因果关系语料标注方法与基准系统；第 4 章提出基于全局优化的因果关系识别方法；第 5 章为实验结果及结果分析；第 6 章为总结与展望。

2 相关工作

Wolff[1]从语言学角度分析人们对因果关系的认知方式，用条件概率之差表示因果关系强度，使用多维特征方法表示因果、促进、抑制三类事件关系。

目前，事件因果关系识别研究大多针对英文语料，研究方法主要可分为模板匹配方法和机器学习方法。

在模板匹配方法方面，Mirza[2]提出了一种事件对显式因果关系标注方法，并综合考虑事件事实性、极性等属性，对标注结果进行修正。Girju[3]抽取显式因果关系，并使用规则和语言学方法抽取部分隐式因果关系用于问答系统。Christopher[4]利用句法树信息，抽取特定结构的子树，使用文本特征与句法树结构对训练语料得到的模板进行匹配来识别因果关系。Ittoo[5]使用半监督方法，使用少量已知的因果关系模板，寻找新的因果关系，再利用新找到的因果关系，扩展出新的模板，在语料库中循环迭代。Radinsky[6]等收集大量新闻语料，使用因果关系构成事件网络，对于给定事件，找出与其相似的历史事件，利用历史事件的演化发展过程预测该事件的发展趋势。

在机器学习方法方面，Bethard[7]等将因果关系与时序关系相关联，提取事件之间文本特征与时序关系构造分类器模型进行识别。Rink[8]使用文本特征和最大频繁子图方法，识别事件因果关系，其使用的特征中，包括人工标注的事件时序关系特征，实验结果表明时序关系特征能够大大提高因果关系识别性能。Do[9]使用多种相似度度量方法，计算事件之间相似度，并提出事件之间的基本约束条件，识别事件因果关系。

由于机器学习方法的事件关系识别结果可能造成逻辑矛盾，因此有少量学者提出使用全局优化方法，构建事件对的时序关系的约束提升识别性能。

Chambers[10]在机器学习方法的基础上，使用整数线性规划方法，使用时序关系的传递性来提升事件时序关系识别性能。郑新[11]使用整数线性规划方法，使用文本信息提出多种事件约束，提升中文事件时序关系识别性能。Denis[12]使用整数线性规划方法，提出共指与回指之间的多种约束，提升共指消解任务的识别性能。本文是第一个在事件因果关系识别的全局优化模型，从语言学的角度对因果关系进行条件约束，从而提高了性能。

3 语料库标注及基准系统

3.1 因果语料库标注

在因果关系研究中，因果关系语料库都是基于英文标注，且主要标注同一句子内的事件关系。为了研究中文事件的因果关系，本文标注一个中文事件因果关系语料。其中，事件因果关系分为显式因果关系和隐式因果关系。

显式因果关系由一个表示因果关系的词连接两个事件，且通常在同一句子内，因此易于

标注，如例 1 所示：

例 1：此外一名日本籍管工***殴打(E1)***工人也是工人们***罢工(E2)***的原因之一。

本文构建一个中文因果标志词词表，如果两个事件在同一句子内，且事件之间连接词在词表中，则标注为因果关系。常用的中文因果标志词如表 1 所示。

表 1：部分中文因果标志词

因为	由于	因此	以致于
于是	致使	导致	因
造成	结果	所以	因而
故	原因	引发	之所以

文章中的事件通常分布在各个句子或段落中。而对于跨句子或者段落的事件，一般不存在明确表示因果关系的词语，需要根据文章内容和语义知识判别因果关系，如例 2 所示：

例 2： 1） 因为机师误闯跑道冲撞护栏，造成机身爆炸起火，目前已经确定有 82 人***死亡(E1)***。

2） 并不排除以“业务过失致死”罪***起诉(E2)***3 位机师。

从上例中看出，因为事故造成人员死亡，所以起诉机师。在两个句子中，没有明确因果关系词连接“死亡”与“起诉”事件，这类关系称为隐式因果关系。

针对隐式因果关系标注，本文参考 Wolff[1]提出的两个因果关系判别模型进行标注，对于两个事件C和E：

1） 当C发生时，E发生的概率远大于C不发生时，E发生的概率，则认为C与E为因果关系，表达式如下：

$$P(E|C) >> P(E|\neg C) \Rightarrow C \rightarrow E \quad (1)$$

2） C和E为因果关系，当且仅当如果C不发生，则E不发生

$$C \rightarrow E \Leftrightarrow \neg C \rightarrow \neg E \quad (2)$$

在标注时，某些事件关系具有歧义，有部分事件之间虽然满足以上定义，但仍标注为非因果关系，总结有如下三类：

1) 两个事件相互对应时，一个事件发生，则另一事件必然会发生。

例 3：他于 22 号凌晨因为胸部痛疼***住进(E1)***了乔治华盛顿大学医院。目前康复良好，很快就可以***出院(E2)***。

2) 两个事件互为共现事件，即两个事件经常共同出现。

例 4：这起冲突结果造成两名巴勒斯坦人***丧生(E1)***和 10 多人***受伤(E2)***。

3) 两个事件为目的关系。

例 5：北韩政府高级官员***访问(E1)***华盛顿的前期，美国和北韩就国际恐怖主义进行了***会谈(E2)***。

3.2 语料库标注结果

本文使用 ACE2005 中文语料库作为基础语料，该语料按照来源由三部分组成，分别为 broadcast news、newswire 和 weblog。为增大因果关系比例，需用事件之间联系较大的语料，因此本文选取其中来源为 broadcast news 的文档作为实验语料。语料共包含 298 篇文档，1398 个事件实例，按文档组成事件对，共 9300 个事件对。标注结果如表 2 所示。其中，显式因果关系和隐式因果关系数量如表 3 所示。从表 3 可以发现，文中的隐式因果关系比例达到 93.02%。这充分说明了在因果关系识别中识别隐式因果关系的重要性。如何识别这些隐式因果关系，是本文研究的重点。很明显，充分利用各种篇章知识和语义信息，是识别隐式因果关系的关键。

表 2 ：语料标注结果

	因果关系	非因果关系
数量	2206	7094
所占比例	23.72%	76.28%

表 3：显式、隐式因果关系比例

	显式因果关系	隐式因果关系
数量	154	2052
所占比例	6.98%	93.02%

3.3 基准系统

本文参考 Bethard[7]和 Rink[8]等使用的特征，结合黄一龙[17]识别中文事件相关性方面使用的特征，作为本文的基准系统。使用的特征如下表 4 所示。

表 4：因果关系识别特征

特征类别	具体特征
词汇特征	事件触发词
	触发词左右窗口词性
	触发词词性
	触发词相似度
	事件之间因果词
事件特征	事件属性（类型、形态、极性、泛型、时态）
	是否存在相同论元
	事件地点是否相同
句法特征	事件所在句子距离
	事件对句法树路径
	事件对依存树路径

4 全局优化模型

分类器独立地对每一对事件进行因果关系识别，而没有利用文档内其他事件对的信息，识别结果容易产生矛盾。为此，本文提出使用整数线性规划的方法进行事件因果关系识别，特别是隐式因果关系识别。该方法结合多种约束条件，达到事件对概率之和最大化。

4.1 目标函数

本文以文档为单位，对文档内的所有事件对，提出以下优化目标函数：

$$\arg\max_{x} \sum_{e_i,e_j \in E} \sum_{r \in \{c,\bar{c}\}} x(e_i,e_j,r) \cdot \log P(r \mid e_i,e_j) \qquad (3)$$

其中，E 为文档内所有事件的集合，r 为事件因果关系，取值为 c 时表示具有因果关系，取值为 $\bar{c}$ 时表示没有因果关系。$x(e_i,e_j,r) \in \{0,1\}$，取值为 1 时，表示事件对具有 r 关系，否则表示没有 r 关系。$P(r \mid e_i,e_j)$ 表示分类器（即 3.3 所述的基准系统）得到的事件对之间 r 关系的概率。当没有任何约束时，最优化目标函数等价于完全使用分类器分类结果。但是，分类器没有考虑同文档内其他事件的信息，为此，本文提出一系列约束，根据文档信息确定部分事件对关系，使目标函数最优化。

4.2 基本约束条件

基本约束条件是文档内事件对之间必须满足的约束条件。本文提出四个基本约束条件，分别是唯一性、非自反性、同指传递性、非传递性。

唯一性：事件对(e_i,e_j)非之间的关系是唯一的，即因果或非因果关系：

$$x(e_i,e_j,c)+x(e_i,e_j,\bar{c})=1 \ \forall e_i,e_j \in E \tag{4}$$

非自反性：如果对于事件对(e_i,e_j)，如果其具有因果关系$x(e_i,e_j,c)=1$，则$x(e_j,e_i,c)=0$，表达式描述如下：

$$x(e_i,e_j,c)+x(e_j,e_i,c)\le 1 \quad \forall e_i,e_j \in E \tag{5}$$

同指传递性：对于两个事件对(e_i,e_j)和(e_i,e_k)，如果$x(e_i,e_j,c)=1$且e_j与e_k为同指事件，则$x(e_i,e_k,c)=1$，表达式描述如下：

$$x(e_i,e_j,c)=1 \wedge coreferenœ(e_j,e_k) \Rightarrow x(e_i,e_k,c)=1 \ \forall e_i,e_j,e_k \in E \tag{6}$$

非传递性：对于三个互相不同指的事件e_i,e_j,e_k，如果$e_i \to e_j$且$e_i \to e_k$，则$e_j \not\to e_k$，表达式描述如下：

$$x(e_i,e_j,c)+x(e_i,e_k,c)+x(e_j,e_k,c)\le 2 \quad \forall e_i,e_j,e_k \in E \tag{7}$$

4.3 限定性约束条件

4.3.1 因果标志词

如果两个事件在同一句子内，且事件之间存在因果标志词，则将其关系置为因果关系。

$$Dist(e_i,e_j)=1 \wedge conj \in Causal\_Set \Rightarrow x(e_i,e_j,c)=1 \ \forall e_i,e_j \in E \tag{8}$$

4.3.2 事件类型约束

文章表达内容往往以句子为单位，一句话内的事件有强烈的相关性。而由于中文表达方式的特点，经常省略主语等信息，从而无法完整获取事件的论元。因此，提出事件类型约束。描述如下：

如果两个事件在同一句子内，且两个事件类型$type_i$和$type_j$在开发集内的共现次数大于某个阈值T，且两种类型之间为因果关系的比率大于某个阈值α_1，则将其关系置为因果关系。

$$\begin{aligned} &Dis(e_i,e_j)=1 \wedge Cou(type_i,type_j) \ge T \\ &\wedge Rat(cr(c \mid type_i,type_j)) \ge \alpha_1 \Rightarrow x(e_i,e_j,c)=1 \ \forall e_i,e_j \in E \end{aligned} \tag{9}$$

4.3.3 论元角色约束

事件发生会涉及若干论元，而如果两个事件有相同论元，则它们之间为因果关系的可能性更大，如例 6 所示：

例 6： 他被***控(E1)***于 1989 年同其他几名成员一道将一名试图脱离这个组织的 21 岁的成员***谋杀(E2)***。

在上例中，“他”是事件“控”与“谋杀”事件的论元之一，且这两个事件之间为因果关系。但是，也有大量有相同论元的事件对之间为非因果关系，如例 7 所示：

例 7：行政院长唐飞上午 7 点多钟就***抵达(E1)***行政院，召开财经***会谈(E2)***。

每个论元在事件中都有其特定的角色，当论元角色是“攻击者”、“受害者”者等与语境有强烈相关性的角色时，我们认为该论元是重要论元角色。本文构建重要论元角色列表，构造方法如下：

在训练集中对每个事件，抽取其论元所属的论元角色，统计每类论元角色在因果关系中的比率α_2，当大于某一阈值时，认为其为关键论元角色。最终构建重要论元角色列表Arg_Set。

因此，提出论元角色约束：如果第一个事件中的关键论元角色的同指论元在事件二所在句子中出现，则认为事件对之间为因果关系，具体表示为：

$$\exists \arg_k \in e_i \wedge \arg_k \in Arg\_Set \wedge Dist(\arg'_k, e_j) = 1 \Rightarrow x(e_i, e_j, c) = 1 \ \forall e_i, e_j \in E \quad (10)$$

5 实验结果

5.1 实验设置

本文使用 ICTCLAS2015[2]工具进行分词，Stanford Parser[3]进行句法分析和依存分析。使用 Mallet[4]工具包中的最大熵分类器，使用 Gurobi[5]工具进行全局优化，按照文档将事件组成事件对，进行 5 倍交叉验证。使用准确率(Accuracy)、召回率(Recall)和 F1 值作为系统的性能评价指标。

5.2 实验结果及分析

经开发集调试，实验参数事件类型对出现次数T取 10，阈值α_1取 0.8，重要论元角色阈值α_2取 0.2。表 5 给出了因果关系识别的结果，本文方法比基准系统在 F1 值上提高了 5.54%。这充分说明了本文提出的全局优化方法在识别事件因果关系，特别是隐式因果关系方面的作用。另外，表 6 列出了各个约束条件组合后的贡献。

表 5：因果关系识别结果

	Precision(%)	Recall(%)	F1(%)
基准系统	59.71	52.14	55.67
+基本约束	60.91	52.13	56.18(+0.51)
+基本约束+标志词	59.90	55.94	57.85(+2.18)
+基本约束+事件类型	62.82	53.49	57.59(+1.92)
+基本约束+论元角色	62.26	53.40	57.49(+1.82)
总体性能	60.94	61.47	61.21(+5.54)

表 6：各约束贡献度

	Precision(%)	Recall(%)	F1(%)
基本约束条件	60.91	52.13	56.18
+标志词+事件类型	60.62	58.48	59.53(+3.35)
+标志词+论元角色	60.27	58.93	59.59(+3.41)
+事件类型+论元角色	63.17	56.75	59.75(+3.57)
+所有约束	60.94	61.47	61.21(+5.03)

从表 5 可看出，本文使用的四种约束均为有效约束，在基准系统上性能都有所提高。基本约束条件能够保证分类结果之间没有歧义。其中，同指传递性能够跨句子或者段落进行信息传递：假设两个事件所处位置不在同一句子内，且存在该事件对的两个同指事件，它们在同一句子内，由于同一句子内识别性能较高，可以提升总体性能。

因果标志词显式地表示句子内事件之间的因果关系，对性能提升也最大。而如果简单地考虑两个事件之间的因果标志词，会引入大量噪声，考虑如下的句子顺序结构：$e1, Causal\_Signal, e2, ..., e3$，很显然，因果标志词连接的是事件$e_1$和$e_2$。因此，需要提出结合句法树信息，具有特定句法树结构的事件对之间限定为因果关系，结构如图 1 所示。

[2] http://ictdas.nlpir.org/downloads
[3] http://nlp.stanford.edu/software/lex-parser.shtml
[4] http://mallet.cs.umass.edu/
[5] http://www.gurobi.com/

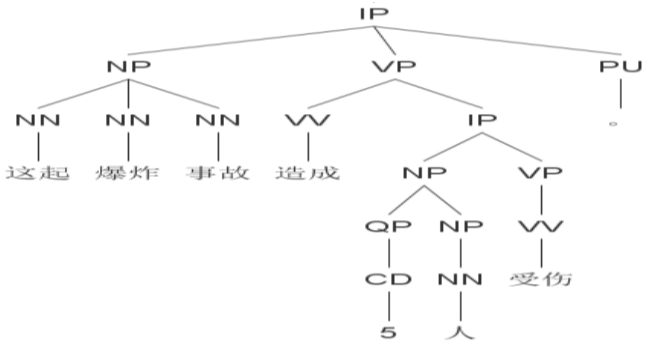

图 1：因果关系句法结构

加入事件类型约束后，准确率和召回率都有所提升，且准确率提升较大。这一结果与预期相符，即有部分事件对之间因论元信息的缺省而造成系统无法识别。论元角色信息约束是针对跨句子和跨段落的事件关系识别的。在不同句子内事件联系较少，且有相同论元的事件较少。但是事件论元往往在另一事件所在的句子中作为实体出现，因此关键论元信息是连接两个句子的重要线索。实验结果表明加入该特征能有效提升系统识别性能。

6 总结

本文提出一种基于全局优化的中文事件因果关系识别方法，其实验结果表明本文提出的方法性能比基准系统有一定提升。本文提出的基本约束条件能够有效消除单纯使用分类器识别造成的结果矛盾。而限定性约束条件能够利用事件对之外的其他事件信息，有效提升系统识别性能。

下一步工作中，可以考虑更多有效的事件之间的特征，提升基准系统性能。其次，考虑更多有效的事件之间的联系，引入更多语义信息提升系统性能，使用全局优化方法进一步提升系统性能。

参考文献

[1] Wolff P. Representing causation[J]. Journal of experimental psychology: General, 2007, 136(1): 82-111.

[2] 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.

[3] Girju R. Automatic Detection of Causal Relations for Question Answering[C]// The Workshop on in the 41 St Meeting of the Association for Computational Linguistics. 2003.

[4] Khoo C S G, Chan S, Niu Y. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns[C]// Meeting of the Association for Computational Linguistics. 2002:336--343.

[5] 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.

[6] Radinsky K, Davidovich S, Markovitch S. Learning causality for news events prediction[C]// International Conference on World Wide Web. ACM, 2012:909-918.

[7] Bethard S, Corvey W, Klingenstein S, et al. Building a Corpus of Temporal-Causal Structure.[C]// International Conference on Language Resources and Evaluation, Lrec 2008, 26 May - 1 June 2008, Marrakech, Morocco. 2008:908-915.

[8] Rink B, Bejan C A, Harabagiu S M. Learning Textual Graph Patterns to Detect Causal Event

Relations.[C]// International Florida Artificial Intelligence Research Society Conference. 2010.
[9] Do Q X, Chan Y S, Roth D. Minimally supervised event causality identification[C]// Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011:294-303.
[10] Chambers N, Dan J. Jointly Combining Implicit Constraints Improves Temporal Ordering[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008.
[11] 郑新. 中文事件时序关系识别与推理方法研究[D]. 苏州大学, 2015.
[12] Denis P, Baldridge J. Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming.[C] // Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, April 22-27, 2007.
[13] Girju R, Dan M. Mining Answers for Causation Questions[C]// AAAI Symposium on. 2002.
[14] The Penn Discourse Treebank 2.0 Annotation Manual, 2007
[15] Bethard, Steven, Martin, James H. Learning semantic links from a corpus of parallel temporal and causal relations[C]// ACL 2008, Proceedings of the, Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, Usa, Short Papers. 2008:177-180.
[16] Hashimoto C, Torisawa K, Kloetzer J, et al. Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features[C]// The, Meeting of the Association for Computational Linguistics. 2014.
[17] 黄一龙, 李培峰, 朱巧明. 中文事件相关性语料库构建及识别方法[J]. 计算机工程与科学, 2015, 37(12):2306-2311.

作者联系方式：

姓名：黄一龙　地址:江苏省苏州市十梓街 1 号　邮编：215006
电话：18306210120　电子邮箱：yilonghuang123@163.com

姓名：李培峰　地址:江苏省苏州市十梓街 1 号　邮编：215006
电子邮箱：pfli@suda.edu.cn

姓名：朱巧明　地址:江苏省苏州市十梓街 1 号　邮编：215006
电子邮箱：qmzhu@suda.edu.cn