基于Self-Attention的因果关系提取具有转移嵌入的 BiLSTM-CRF

李兆宁、李琪、邹晓天、任江涛*

中山大学数据与计算机科学学院, 广东省广州市 510006

抽象的

从自然语言文本中提取因果关系是一个具有挑战性的开放问题

莱姆在人工智能中。现有的方法利用模式、约束和

机器学习技术来提取因果关系,很大程度上取决于领域

知识并且需要大量的人力和时间来进行特征工程

尼宁。在本文中,我们将因果关系提取表述为序列标记

基于新颖的因果关系标记方案的问题。在此基础上,我们提出

以 BiLSTM-CRF 模型为骨干的神经因果关系提取器,

命名为SCITE （Self-attentive BiLSTM-CIth Transferred Embeddings ）,可以直接提取因果关系,而不需

要提取候选因果关系

配对并分别确定它们的关系。解决数据问题

不足之处,我们转移上下文字符串嵌入,也称为 Flair

嵌入,在我们的任务中在大型语料库上进行训练。此外,

为了提高因果关系提取的性能,我们引入了多头

将自注意力机制引入 SCITE 中,以学习因果关系之间的依赖关系

字。我们在公共数据集和实验结果上评估我们的方法

证明我们的方法取得了显着且持续的改进

与基线相比。

关键词:因果关系提取、序列标记、BiLSTM-CRF、Flair

*通讯作者 邮箱地址:
lizhn7@mail2.sysu.edu.cn（李兆宁）、liqi38@mail2.sysu.edu.cn（李琪）、zouxt5@mail2.sysu.edu.cn（邹小天）、issrjt@ mail.sysu.edu.cn (任江涛)

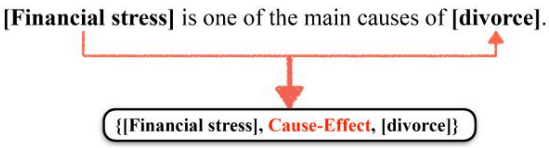


图1:表达因果关系的句子,在本例中,“financialstress”是原因,
“离婚”是“经济压力”造成的影响。

嵌入,自注意力

一、简介

自然语言文本包含相当多的因果知识,如图所示

图1.近年来,因果关系提取变得越来越重要

对于许多自然语言处理任务,例如信息检索 [1, 2]、事件预测 [3, 4]、问答 [5, 6, 7]、生成未来场景 [8, 9]、决策处理 [10],医学文本挖掘[11,12,13]和行为预测[14]。然而,由于自然语言的歧义性和多样性

文本中,因果关系提取仍然是一个很难解决的 NLP 问题。

传统的因果关系提取方法可以分为两类:

gories:基于模式的方法[1,11,15,16]（第5.1节）,以及基于模式和机器学习技术组合的方法[5,17,18,19]（第5.2节）。前者往往跨域适用性较差,失败

平衡精确度和召回率,可能需要广泛的领域知识

解决特定领域的问题。后者通常需要相当大的
特征工程上的人力和时间,严重依赖手册
文本特征的选择。一般来说,它把因果关系提取分为两种
子任务、候选因果对提取和关系分类（过滤
非因果对）。候选因果对提取的结果可能会影响
关系分类的性能并产生级联错误。

郑等人。 [20]首先提出了一种标记方案,使得
同时提取实体和关系的模型。在他们的标签方案中,
他们应用实体提及标签和关系类型的笛卡尔积
标签,然后分配一个唯一的标签来编码实体提及和关系
每个单词的类型。受到他们新颖想法的启发,我们专注于因果三元组
它由两个事件实体及其关系组成。例如,
图 1 中的句子包含一个因果三元组:“{财务压力、因果关系、离婚}”。因此,我们可以直
接对因果三元组进行建模,而不是破坏
因果关系提取为两个子任务。根据动机,我们制定
将因果关系提取到序列标记问题中并提出因果关系
标签方案（第2.1节)以实现直接因果关系提取。然而,Zheng等人提出的标记方案。 [20] 无法识
别重叠
句子中的关系;它只考虑实体所属的情况
到一个三元组:如果一个实体参与多个关系,它的标签应该
不是独一无二的。为了解决这个问题,我们设计了tag2triplet算法
（第 2.2 节)处理多个因果三元组和嵌入因果三元组
同一句话。最后,我们将因果关系标记方案与
深度学习架构（第 2.3 节)可最大限度地减少特征工程,同时
有效地建模自然语言文本中的因果关系。

我们注意到一些研究人员还提出了深度学习技术
近年来基于因果关系提取的方法（第 5.3 节）。尽管他们的作品值得称赞,但有些作品
[21,22,23,24]只是一个分类
因果关系而不是提取完整的因果三元组,并且
其他人[25, 26]主要关注语言表达的识别
本文中的因果关系而不是常识性的因果关系提取。

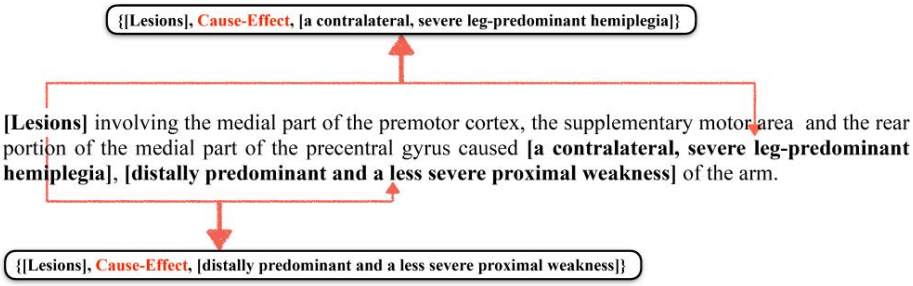


图 2:第二个因果三元组： “{[病变]、因果关系、[远端占主导地位和不太严重的近端弱点]}”几乎涵盖整个句子。

通过应用我们的因果关系标记方案,我们使用基于 BiLSTM 的模型-CRF [27] 直接提取因果三元组。但我们发现有两个障碍阻碍了深度学习模型性能的进一步提升。

首先,在没有任何先验知识的情况下很难训练出优秀的深度学习模型现有语料库数据不足的情况下的知识[28,29,30]。为了缓解这个问题,我们将 Flair 嵌入 [31] 纳入我们的任务中，它使用在大型数据集上训练的字符语言模型的内部状态语料库来创建词嵌入（第 2.3.2 节）。实验结果表明这种上下文字符串嵌入引发了一种新技术趋势在 NLP 中可以极大地提高因果关系提取的性能。

其次,就其在文本中的位置而言,因果关系是一些时间彼此相差很远,如图2所示。中的长程依赖因果三元组在深度学习模型中造成了困难和模糊性,但是一组基于依赖树的逻辑规则可以轻松准确地执行牵引这样的三胞胎。了解原因之间的这种远程依赖性和效果,我们将多头自注意力机制[32]引入到我们的模型中（第2.3.4节）。与递归处理的基于 LSTM 的模型不同每个单词之间,自注意力机制可以进行直接连接句子中的两个任意单词,从而允许畅通无阻的信息流通过网络[33]。

本文的贡献可概括如下：

1.我们设计了一种新颖的因果关系标记方案来直接提取因果关系在文本中,可以轻松地将因果关系提取转换为序列标记任务并处理多个因果三元组和嵌入因果同一句话中的三连音。

2.基于我们的因果关系标记方案,我们提出了SCITE （Self-attentive BiLSTM-CRF with Transferred Embeddings）,一种基于神经的因果关系具有在大型数据集上训练的转移上下文字符串嵌入的提取器语料库。据我们所知,我们是第一个传递天赋的人嵌入到因果关系提取中。

3.我们将多头自注意力机制引入到SCITE中，使模型能够捕获原因和原因之间的远程依赖关系影响。

4. 广泛的实验结果（第 3 节)和进一步分析（第 4 节）表明我们的方法取得了显着且一致的改进与其他基线相比。我们将代码和数据集发布到研究社区进行进一步研究[1]。

2. 方法

2.1.因果关系标签方案

我们用“BIO”（begin、inside、other)和“C、E、Emb”（cause、effect、embedded causality)符号来表示单词的位置信息以及因果事件的语义角色,分别嵌入因果关系[30]表明因果事件在因果关系中具有不同的作用不同的三胞胎。图 3 是句子中嵌入因果关系的示例。该例句包含两个因果三元组：“{慢性炎症、因果、产酸增加}”和“{螺杆菌、因果、慢性炎症}”,请注意“慢性炎症 - “mation”是第一个三元组中的因,第二个三元组中的果。

1https://github.com/Das-Boot/scite

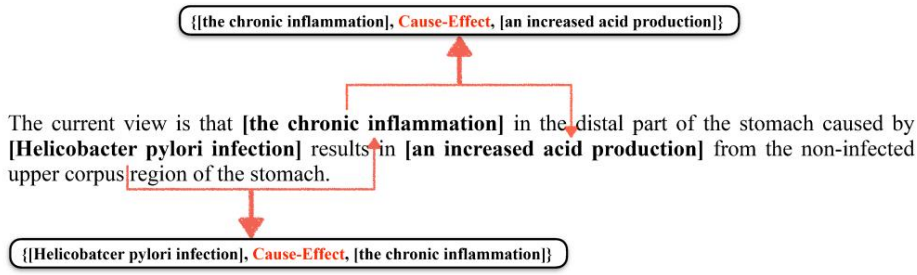


图 3:两个因果三元组共享相同的因果事件实体:“慢性炎症-化”在一个句子中。

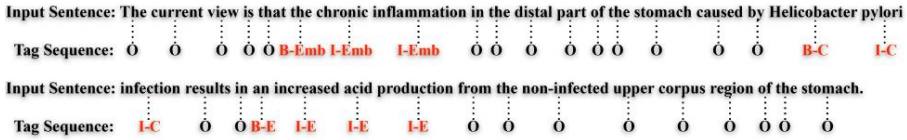


图 4:基于我们的因果关系标记的例句的标准注释方案。

图 4 显示了这种因果关系序列标记的示例。基于我们的因果关系标记方案,我们将因果事件实体标记为“慢性炎症”化”、“幽门螺杆菌感染”和“产酸增加”分开带有我们的特殊标签。具体来说,标签“O”代表“其他”,即意味着相应的词在任何因果关系成分中都是不相关的。标签“BC”代表“原因开始”,标签“IC”代表“原因发生”side”,标签“BE”代表“效果开始”,标签“IE”代表“效果”inside”,标签“B-Emb”代表“嵌入因果关系开始”,标签“I-Emb”代表“内在的因果关系”。因此,总数量为标签为Nt = 7。

2.2.从标签序列到因果三元组

我们设计了一个tag2triplet算法来自动获得最终的从图 4 中的标签序列中提取了三元组。为了更好地说明该算法,rithm,我们定义两种类型的因果关系:简单因果关系和复杂因果关系。

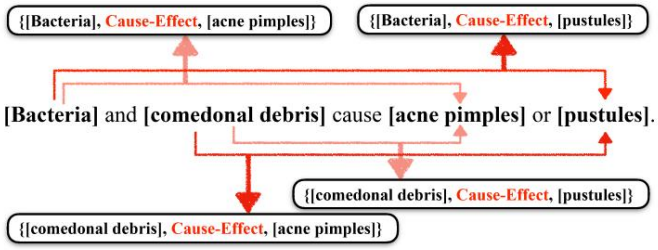


图 5:对于上述句子中的四个因果三元组中的任何一个,都有另一个因果关系
三联体具有相同的原因或结果。

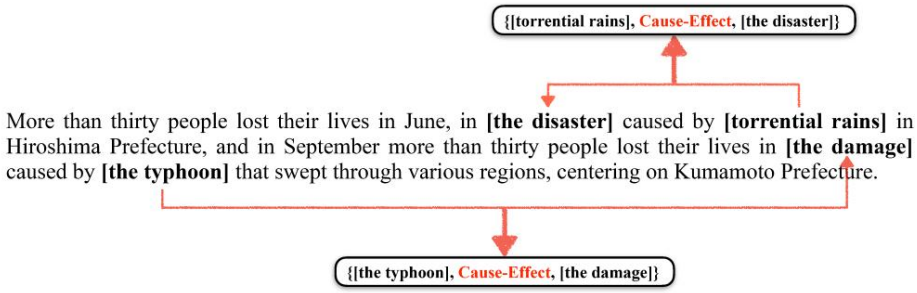


图6:因果三元组“{暴雨、因果、灾害}”和“{台风、因果、损害}”不具有相同的因果关系。

2.2.1.简单因果关系的例子

简单因果关系可以分为两类:

1. 句子中只有一个原因或一个结果,并且没有
   嵌入因果关系,即NC = 1 或NE = 1 且NEmb = 0,其中
   数控、 NE、 NEmb分别表示标签“BC”、“B-”的数量
   E”,以及句子中的“B-Emb”。图1和图1中的例句
   图2 都是这种类型的因果关系。
2. 句子中有多个原因和结果,并且没有
   嵌入因果关系,即NC > 1、 NE > 1 且NEmb = 0。此外,对于
   句子中的每个因果三元组,必须至少有一个因果三元组
   具有相同的原因或结果。图5中的例句是
   这种类型的因果关系。

2.2.2.复杂因果关系的案例

复杂因果关系有以下两种类型：

1. 句子中存在嵌入因果关系，即NC > 0, NE > 0 且 NEmb > 0。图3中的例句就是这种类型的因果关系。
2. 句子中有多个原因和结果，并且没有嵌入因果关系，即NC > 1、 NE > 1 且NEmb = 0。此外，在句子中的所有因果三元组，必须至少有一个因果与任何其他三元组不具有相同原因或结果的三元组。图6中的例句就是这种因果关系。请注意，图5句子中的因果关系分布与图6：前者中的每个因果三元组混合在一起，并且每个因果三元组混合在一起后者的三联体是分开的。

2.2.3. Tag2三元组算法

tag2triplet 算法在算法 1 中描述。我们详细阐述 tag2triplet 算法，通过获取句子 S 及其对应的标签图4以序列Stag为例；中间结果显示在表格1。

首先，我们统计因果关系的出度和入度，并找到索引雄鹿的因果关系。具体来说，“原因”的出度记为1，“effect”的入度记为1，“effect”的出度和入度“嵌入因果关系”都记为1。然后，我们判断是否S根据数量和数量是简单因果关系还是复杂因果关系每个因果标签的分布：“C”、“E”和“Emb”。在Stag 中， NEmb = 1，并且因此，S 是复杂的因果关系。然后，我们应用笛卡尔积由因果标签组成的因果实体，生成因果的候选者三联体。在表1中，候选“(E0, E2)”代表因果三元组“{the 慢性炎症、因果关系、产酸增加}。

接下来，组合从输入返回 i 长度的三元组子序列候选人。之后，我们判断是否是出度和入度

```
input :句子S对应的标签序列Stag
输出:句子 S 中的因果三元组

计算 Stag 中因果关系的出度和入度;
找到Stag中因果关系的索引idx ;
如果 S 中的因果关系 ∈ 简单因果关系 那么
    候选 ← 笛卡尔积(idx);
    如果 CheckConjunction(candidate, idx, S) 为 true 那么
        因果三元组←候选;
结束
结束
如果 S 中的因果关系 ∈ 复杂因果关系 那么
    候选人 ← 笛卡尔积(idx);
    for i ← Max(Sum(出度), Sum(入度)) 至
      Len（候选人)做
        标志←0;
        记录 ← [ ];
        对于 j ∈ Combination(candidates, i) 做
            如果 CheckDegree(j, 出度, 入度) 为 true 并且
              CheckConjunction(j, idx, S) 为 true 那么
                距离 ← SumDistance(j, idx);
                AppendToRecord(j, 距离);
                标志 ← 1;
            结尾
        结尾
        如果标志 = 0 那么
            休息;
        结尾
结束
    因果三元组 ← Min(records, key=records[－1]);
结束
```

算法 1:Tag2triplet

表1:当我们输入句子S及其时,运行tag2triplet的中间结果
对应的标签序列Stag,我们突出显示候选者的正确组合
大胆的。

S	... [慢性炎症]E0 ... [幽门螺杆菌感染]E1 ... [产酸增加]E2 ...		
麈	[B-Emb I-Emb I-Emb]E0	[BC IC IC]E1	[BE IE IE IE]E2
指数	[5,6,7]E0	[17,18,19]E1	[22,23,24,25]E2
出度	1E0	1E1	0E2
入度	1E0	0E1	1E2
候选人	(E0, E2)	（E1、 E0）	（E1、 E2）
组合	(E0, E2), (E1, E0)	（E0， E2），（E1， E2）	(E1, E0), (E1, E2)
出度	（1E0、1E1、0E2） _ _ _	（1E0、1E1、0E2） _ _ _	（1E0、0E1、0E2） _ _ _
入度	（1E0、0E1、1E2） _ _ _	（0E0、0E1、1E2） _ _ _	（1E0、0E1、1E2） _ _ _

每个候选组合都与原来的出度一致并且
雄鹿的入度。然后,我们判断该组合是否匹配
根据 S 中的并列连词进行规则,例如,如果有
同一子句中相邻原因之间的并列连词“and”，
那么这两个原因将形成各自的因果三元组,具有相同的效果，
如图5中的“细菌”和“粉刺碎片”。最后,我们选择
与通过的组合距离最短的组合
检查作为提取的因果三元组。由于只有一个组合“(E0, E2), (E1, E0)”通过了所有检查,因此我们直接将其输出作为最终结果。

2.3.赛特

图 7 给出了因果序列 SCITE 模型的主要结构
标签。我们取输入句子 S = {xt} n序列 y = {yi} n t=1 及其对应的标签
t=1 为例介绍SCITE的各个组成部分
从下到上如下,其中n是S的长度。

2.3.1.用于字符表示的 CNN

为了捕获特定于任务的子字特征,我们采用相同的卷积
神经网络 [34] (CNN) 架构如 Ma 和 Hovy [35],使用一层

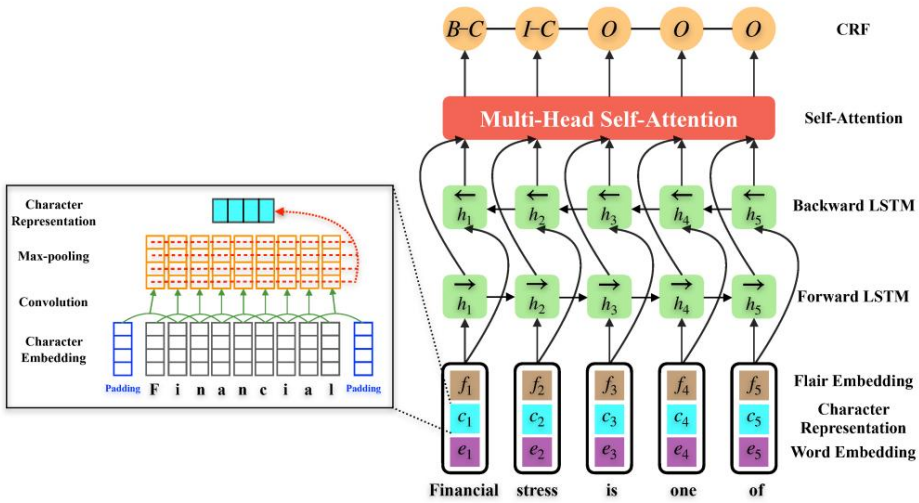


图 7:因果序列标记 SCITE 的主要结构。的左侧
图中显示了代表“金融”一词的字符CNN结构。

CNN 结构后跟最大时间池化操作 [36] 来学习
字符级表示。该过程如图 7 左侧所示。

具体来说,令ri ∈ R m为 m 维特征向量,对应-
查找单词xt中的第 i 个字符（ xt的长度为 s）。卷积运算涉及过滤器 w ∈ R lm,该过滤器
应用于 l 个字符的窗口
产生一个新特征。例如,特征ci是从窗口生成的
字符ri:i+l－1 [2] 经过

$$ci = wTri:i+l-1 + b, \tag{1}$$

其中 b 是偏置项。该过滤器应用于每个可能的窗口
单词{r1:l , r2:h+1, ..., rs－l+1:s}中的字符向量以产生特征
地图

$$c^ = [c1, c2, ..., cs-l+1], \tag{2}$$

[2] 一般来说,让ri:i+j指的是字符向量ri, ri+1, ..., ri+j的串联

其中 $c \in R^{s-l+1}$。然后,我们取最大值 $c = \max\{c\}$ 作为与该特定过滤器相对应的功能。因此,表示该数过滤器的数量为 f,单词xt的字符表示ct给出为:

$$ct = [\tilde{c}_1, \tilde{c}_2, \ldots, \tilde{c}_f] \tag{3}$$

2.3.2.迁移从大型语料库中学到的情境化表示

近年来,深度学习在自然科学领域取得了令人难以置信的进步语言处理(NLP)任务得益于其强大的表示学习能力。但在现有语料库数据不足的情况下,深度学习的饥饿本质限制了我们基于神经的模型的性能在因果关系提取中。最近在大型语料库上训练的情境化语言表示模型 [37,38,31] 的发展揭示了迁移学习的可能性。

在本文中,我们使用迁移学习来缓解数据输入问题充足性。具体来说,我们建议转移 Flair 嵌入[31],它源自在 1- 上训练的字符级语言模型(CharLM)十亿字基准语料库 [39] 来完成我们的任务。这个 CharLM 包含一个前向语言模型(fLM)和后向语言模型(bLM)。跟随-ing Akbik 等人。 [31],我们提取输出隐藏状态 $\overrightarrow{h}^{t}_{end+1}$ 来自fLM在最后一个字符 r结束之后 xt这个词的。同样,我们获得输出 hid- $\overleftarrow{h}^{t}_{start-1}$ 来自第一个字符 r 之前的 bLM t状态 $\overleftarrow{h}^{t}_{start}$ xt这个词的。然后,将两个输出隐藏状态连接起来形成最终的嵌入 F_t^{charlm} 单词xt如下:

$$F_t^{charlm} = [\overrightarrow{h}^{t}_{end+1}, \overleftarrow{h}^{t}_{start-1}] \tag{4}$$

最后,我们连接转移的 Flair 嵌入 f_t^{CharLM} 和 char-角色表示ct与单词嵌入并由Komninos 进行预训练和 Manandhar [40] 并将它们输入 BiLSTM 层。

2.3.3.双向LSTM

长短期记忆（LSTM)[41]是一种特殊的循环神经网络
克服梯度消失和爆炸问题的工作（RNN)[42]
传统的 RNN 模型。通过专门设计的闸门结构
LSTM,模型可以选择性地保存上下文信息。基本单位为
LSTM架构是一个记忆块,其中包括一个记忆单元（de-
记为m)和三个自适应乘法门（即输入门i、忘记门f和输出门o）。形式上,计算操作

在时间 t 更新 LSTM 单元是：

$$it = \sigma(Wi\ et, ct, f_t^{查尔姆} + Uiht-1 + bi), \quad (5)$$

$$ft = \sigma(Wf\ et, ct, f_t^{查尔姆} + Ufht-1 + bf), \quad (6)$$

$$ot = \sigma(Wo\ et, ct, f_t^{查尔姆} + Uoht-1 + bo), \quad (7)$$

$$mt = \tanh(Wm\ et, mt, f_t^{查尔姆} + Umht-1 + bm), \quad (8)$$

$$mt = it\ mt + ft\ mt-1, \quad (9)$$

$$ht = ot \tanh(mt), \quad (10)$$

其中[et, ct, $f_t^{查尔姆}$] 和ht表示输入向量和隐藏状态，
分别在时间t。 σ 是元素级 sigmoid 函数,并且是
元素乘积。 Wi , Wf , Wo, Wm是输入的权重矩阵
矢量,用户界面 、 Uf 、 Uo、 Um是隐藏状态的权重矩阵， bi ,
bf 、 bo 、 bm表示偏置向量。

然而,LSTM只考虑过去的信息,忽略未来的信息
真实信息。为了有效地使用上下文信息,我们可以使用双向
LSTM（BiLSTM）。 BiLSTM 使用前向 LSTM 和后向 LSTM
对于每个序列获得两个单独的隐藏状态：−→ht， ←− ht,然后
时间 t 的最终输出是通过连接这两个隐藏状态形成的：

$$ht = [-\!\!\rightarrow ht, \leftarrow\!\!- ht] \quad (11)$$

因此,BiLSTM层的最终输出对于输入句子S

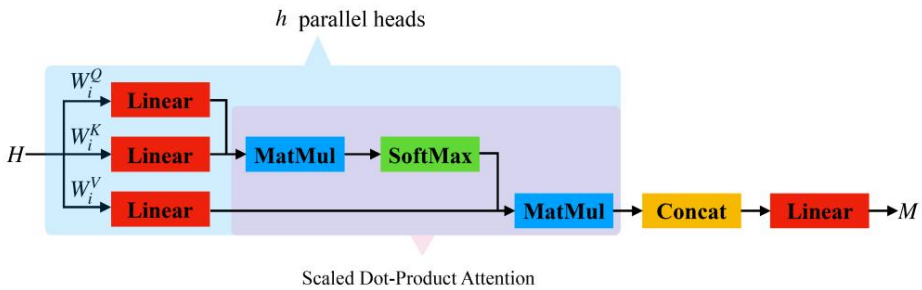


图 8:多头注意力机制的架构。

可以表示为 H = {ht} n t=1，其中 H ∈ R n×d，d 是层大小 BiLSTM 层。

2.3.4.多头自注意力

自注意力是注意力机制的一个特例,它只需要单个序列来计算其表示,已成功应用于许多 NLP 任务 [43,44,32] 并显示了其在捕获方面的优越性ing 远程依赖。在SCITE中,我们采用多头自注意力机制（MHSA)由 Vaswani 等人提出。 [32]了解因果关系的依赖性在给定的句子中。图 8 描述了多头注意力的架构机制。

具体来说,给定 H 作为 BiLSTM 层的输出,多头在 - 注意力机制首先使用不同的学习线性投影将矩阵 H h 次投影到矩阵： HWQ头数和参数矩阵WQ ∈ R d×dv 我， HWK和 HWV i 。其中 h 是 ， WK ∈ R d×dv且 WV ∈ R d×dv是第 i 个头的投影。那么,注意力函数为并行执行,产生n×dv 维输出值。最后,所有的由并行头产生的矩阵被连接起来,产生最终的值M 的维度为 n×(hdv),其中 h 和dv都是超参数自注意力层。配方可表示如下：

$$M = \text{MultiHead}(H, H, H) = \text{Concat}(\text{head}1..., \text{head}h) \tag{12}$$

$$\text{其中 head}i = \text{注意力}(HWQ_{我,}\ \text{华威克}_{我,}\ HWV)_{我} \tag{13}$$

这里,注意力函数是“缩放点积注意力”,即

计算注意力分数如下：

$$\text{注意}(HWQ_{我,}\ \text{华威克}_{我,}\ HWV)_{我} = \text{软 tmax}\left(\frac{(HWQ_{我})(HWK)_{我}^{时间}}{\sqrt{d}}\right)(HWV)_{我} \tag{14}$$

为了充分整合信息,我们将 H 和 M 连接成矩阵 H~,然后用线性投影将 H~ 投影到矩阵:HW~。其中权重矩阵 W ∈ R (d+hdv)k是训练时要学习的模型的参数

k 是不同标签的数量。

2.3.5。病例报告表

条件随机场（CRF)[45]可以获得全局最优链

考虑相邻标签之间的相关性的给定序列的标签。

在序列标记任务中,序列之间通常存在很强的依赖关系

输出标签。因此,不要只使用 RNN 来建模标记决策

分别,我们采用 BiLSTM-CRF [27] 作为 SCIFl 的骨干,共同

解码整个句子的标签。

我们使用 $P \in R^{n\times k}$ 作为线性层输出的分数矩阵,其中

句子中第 i 个单词的第 th 个标签。为了

Pij表示第 j 个句子的得分 $S = \{x_t\}_{t=1}^{n}$ 和标签路径 $y = \{y_i\}_{我=1,}$ CRF给出了实值

得分如下：

$$\text{得分}(S, y) = \sum_{我=0}^{n} \text{阿依,依+1} + \sum_{我=1}^{n} \text{皮,伊,} \tag{15}$$

其中 A 是转移矩阵， Ai,j表示转移的分数

从标签 i 到标签 j。 y0和yn是开头和结尾的特殊标签

一个句子,所以 A 是一个大小为 k + 2 的方阵。因此,概率
对于给定句子 S 的标签序列 y 是:

$$p(y|S)=\frac{e^{\text{分数}(S,y)}}{\sum_{y\in YS} e^{\text{得分}(S,y)}}, \tag{16}$$

我们现在最大化正确标签序列的对数似然:

$$\log(p(y\mid S)) = \text{分数}(S,y) - \log \sum_{y\in YS} e^{\text{分数}(S,y)}, \tag{17}$$

其中YS表示输入句子 S 的所有可能的标签序列。
根据上面的公式,我们可以得到有效的输出序列。解码时,
具有最大得分的序列的输出为:

$$y^{*} = \arg\max_{y\in YS} \text{分数}(S,y) \tag{18}$$

这可以使用动态规划技术来计算,我们
选择维特比算法[46]进行此解码。

3. 实验

3.1.实验设置

3.1.1.数据集

在实验中,我们评估了通过扩展注释获得的语料库
SemEval 2010 任务 8 数据集的选项。 [28]。在原始数据集中,仅
每个句子中的一个因果三元组都被注释。我们扩展注释
SemEval 注释器未考虑因果三元组;例如,
我们注释了图 2 中句子中的所有因果三元组（更多示例如图 3、图 5 和图 6 所示）。具体来说,语料库由
5,236 个句子,其中 1,270 个句子至少包含一个因果三元组。
训练集由 4,450 个句子组成,包含 1,570 个因果三元组。
测试集中有 804 个句子,其中包括 296 个因果三元组。表2
显示数据集六种因果标签的统计信息。

表2:数据集不同类型因果标签的统计

标签类型	训练集	测试集
公元前	1308	236
我知道了	1421	229
是	1268	238
IE	1230	230
B-Emb	55	9
胚胎移植	55	16
和	5337	958

3.1.2.评估

我们使用标准精度（P）、召回率（R)和F1分数（F)作为评估指标,可以通过以下公式计算:

$$P = \frac{\#\text{正确提取因果三元组}}{\#\text{提取的因果三元组}}, \tag{19}$$

$$R = \frac{\#\text{正确提取的因果三元组}}{\#\text{D 中的总因果三元组}}, \tag{20}$$

$$F = 2\frac{\text{普} \cdot R}{\text{普}+R}, \tag{21}$$

其中 D 是数据集中所有句子和预测因果关系的集合当且仅当它与标记的因果关系精确匹配时,三元组才被认为是正确的三联体。为了获得可比较且可重复的 F1 分数,我们遵循以下建议Reimers 和 Gurevych [47] 的研究,每个实验进行 5 次,然后报告平均结果及其标准差,如表 3 所示。

3.1.3.超参数

该模型使用Keras实现[3]版本2.2.4。 300维字采用由 Komninos 和 Manandhar [40] 预训练的嵌入训练过程中保持固定。字符嵌入是随机的

3https://github.com/keras-team/keras

从范围在 [一 $\frac{\sqrt{3}}{\text{暗淡}}$, $\frac{\sqrt{3}}{\text{暗淡}}$],我们在那里

设置 dim = 30。对于字符级 CNN 层,我们使用单层 CNN

30个过滤器,窗口大小为3。我们使用Flair框架 [4] 来计算

天赋嵌入。 LSTM的隐藏大小设置为256。多头自注意力机制的参数h（头的数量)和dv （每个头的大

小)分别设置为3和8。我们使用变分 dropout [48]

以 0.5 的退出率来规范我们的网络。为了解决爆炸问题

梯度问题,我们将阈值为 5.0 的梯度归一化[49]应用于 SCITE。训练过程的优化方法是Nadam [50]

学习率为0.001,我们采用学习率退火方法

如果训练损失在超过 10 个 epoch 内没有下降,则该方法

将使学习率减半。我们让小批量大小为 16。在实验中,

我们对训练集进行网格搜索和 10 倍交叉验证以找到

最佳超参数。在测试集上,我们选择最优模型

在所有 200 个具有最高交叉验证 F1 分数的 epoch 中。

3.1.4.基线

为了进行全面比较,我们将我们的方法与几种方法进行比较

经典的因果关系提取方法,可以分为两类:

基于因果关系标记的管道方法和序列标记模型

方案。我们用作基线的管道方法如下:

- 规则+贝叶斯:Sorgente 等人。 [17]执行模式匹配
  根据一组规则提取候选因果对,然后使用
  贝叶斯分类器和拉普拉斯平滑来过滤非因果对。

- CausalNet:Luo 等人。 [19]建议采用因果强度（CS)来衡量
  任意两段短文本之间的因果强度,整合
  必然因果关系与充分因果关系。为了比较,我们添加
  与 Sorgente 等人相同的因果提取模块。 [17]他们的方法。

4https://github.com/zalandoresearch/flair

然后我们计算候选因果对的CS分数并进行比较

阈值 τ（τ 是一个可调超参数）。如果 CS (c, e) > τ，我们

得出 (c, e) 是因果关系的结论;否则,(c, e) 是错误的

提取对。

本文使用的序列标注结构分为CNN-

基于模型和基于 BiLSTM 的模型。对于基于 CNN 的模型 [51]，

基线如下：

- IDCNN-Softmax:该模型使用深度迭代扩张 CNN（ID-

  CNN)架构来聚合整个文本的上下文,其中有

  比传统 CNN 具有更好的容量和更快的计算速度

  LSTM,然后映射 IDCNN 的输出来预测每个标签

  悬而未决地通过一个softmax分类器。

- IDCNN-CRF:该模型使用CRF分类器来最大化标签

  基于IDCNN的完整句子的概率。相比于

  softmax 分类器,CRF 分类器更适合以下任务

  强输出标签依赖性。

基于 BiLSTM 的模型的基线如下：

- BiLSTM-softmax [52]:模型由两部分组成:BiLSTM

  编码器和 softmax 分类器。

- BiLSTM-CRF [27]:序列标记的经典且流行的选择

  任务,由 BiLSTM 编码器和 CRF 分类器组成。

- CLSTM-BiLSTM-CRF [53]:分层 BiLSTM-CRF 模型

  使用基于字符的表示来隐式捕获形态

  通过字符 LSTM 编码器 (CLSTM) 提取特征（例如前缀和后缀）,然后连接字符嵌入和预训练

  词嵌入作为 BiLSTM-CRF 的输入。

- CCNN-BiLSTM-CRF [35]:类似的分层 BiLSTM-CRF 模型
  使用字符 CNN 编码器 (CCNN) 代替 CLSTM 来学习
  字符级嵌入。

为了进一步分析转移到我们的 Flair 嵌入的性能
任务中,我们结合了 ELMo [37] 和 BERT [38],这两个强大的上下文化模型
单词表示,进入我们特定任务的 BiLSTM-CRF 架构
实验基线:

- ELMo-BiLSTM-CRF:BiLSTM-CRF 的扩展,其中 Peters
  等人。 [37] 将预训练的静态词嵌入与 ELMo（来自语言模型的嵌入)表示连接起来,并将它们视为
  BiLSTM-CRF 的输入。
- BERT-BiLSTM-CRF:Devlin 等人的类似扩展。 [38] 添加了预训练的词嵌入和 BERT（来自
  Transformer 的双向编码器表示)表示,并将它们用作
  BiLSTM-CRF 的输入。
- Flair-BiLSTM-CRF:该模型用作我们的强大基线
  工作,其中 Akbik 等人。 [31] 预训练的词嵌入是串联的
  与 Flair 嵌入结合并将其输入 BiLSTM-CRF 模型。
  请注意,使用 Flair 嵌入的模型已经实现了当前的
  最先进的结果在一系列序列标记任务中,例如命名
  实体识别、分块和词性标注 [31, 54]。
- Flair+CLSTM-BiLSTM-CRF:Akbik 的简单扩展
  等人。 [31] 添加了从任务训练中学习的字符表示
  CLSTM 到 Flair-BiLSTM-CRF。

3.2.实验结果

不同模型在因果关系提取上的性能如图所示
表 3. 第一部分是管道方法（从第 2 行到第 3 行）。第二部分（第4行到第5行)是基于CNN的序列标记方法。这

表 3:测试集上的精度 (P)、召回率 (R) 和 F1 分数 (F) 的比较
基线。

模型	磷	右	F
因果网	0.6211	0.5372	0.5761
规则-贝叶斯	0.6042	0.5878	0.5959
IDCNN-softmax	0.7455±0.0142	0.7074±0.0168	0.7258±0.0105
IDCNN-CRF	0.7442±0.0225	0.7142±0.0122	0.7288±0.0160
BiLSTM-softmax	0.7744±0.0183	0.7622±0.0114	0.7682±0.0138
CLSTM-BiLSTM-CRF	0.8144±0.0284	0.7412±0.0073	0.7757±0.0107
CCNN-BiLSTM-CRF	0.8069±0.0199	0.7520±0.0227	0.7780±0.0075
BiLSTM-CRF	0.7837±0.0061	0.7932±0.0087	0.7884±0.0072
BERT-BiLSTM-CRF	0.8277±0.0058	0.8209±0.0093	0.8243±0.0049
Flair+CLSTM-BiLSTM-CRF	0.8403±0.0090	0.8284±0.0125	0.8343±0.0106
ELMo-BiLSTM-CRF	0.8361±0.0135	0.8399±0.0063	0.8379±0.0092
Flair-BiLSTM-CRF	0.8414±0.0079	0.8351±0.0141	0.8382±0.0092
赛特 (Flair+CCNN-BiLSTM-MHSA-CRF)	0.8333±0.0042	0.8581±0.0021	0.8455±0.0028
赛特 （基于通用标记方案）	0.7609±0.0170	0.7757±0.0136	0.7682±0.0145

第三部分（第6行到第9行)是基于BiLSTM的序列标记方法，
第四部分（第10行到第13行)是使用序列标记方法
上下文化的词嵌入。我们的 SCITE 模型显示在最后部分，
其中第一行是基于建议的标记方案的 SCITE 结果
第二行是基于通用标记方案的SCITE结果。

表 3 显示 SCITE 的 F1 分数优于所有其他模型
测试集中的值为 0.8455。这证明了我们提出的方案的有效性
方法。此外，它还表明序列标记模型更好
比管道方法。

通过比较序列标记模型在测试中的性能
集，我们可以看到基于 BiLSTM 的模型比基于 CNN 的模型更好
楷模。基于 BiLSTM 的模型性能优越的原因可能是
是 LSTM 层可以更有效地捕获全局单词上下文

图 9:我们的 SCITE 模型和其他标签错误基线模型的混淆矩阵。

x轴:真实标签； y 轴:预测标签。

信息并学习因果关系的语义表示。此外,它
还表明,喂食后模型的性能显着提高
将语境化的单词表示放入 BiLSTM-CRF 架构中。在
特别是,Flair-BiLSTM-CRF 实现了 6.32% 的最高改进
与 ELMo 和 BERT 相比,BiLSTM-CRF 的性能优于 BiLSTM-CRF（分别提高了 6.28% 和 4.55%）,这表明上下文化的字符级别
词嵌入更适合因果关系提取的任务。

此外,我们还发现我们提出的因果关系标记方案产生
比一般标记方案（0.8455 与 0.7682)更好的结果
SCITE 架构,验证了我们提出的标记的有效性
方案。一般标记方案不包含“Emb”标记,因此,
该模型无法正确识别嵌入的因果关系。虽然数量
嵌入因果关系在测试集中相对较小,嵌入因果关系
在因果关系提取中起着至关重要的作用:其识别中的一个错误可能会导致
影响多个三元组的正确提取,如图3所示。

四、分析与讨论

4.1.误差分析

在本文中,我们重点关注从自然语言中提取所有因果三元组
标语文本,其中标签“C”（原因）、“E”（结果)的准确识别

表4：“C”（原因）、“E”（结果)和“Emb”预测标签的比较
（嵌入因果关系)测试集上的精度 (P)、召回率 (R) 和 F1 分数 (F)。

模型	CP	CR	CF	EP	ER	EF	Emb-P	Emb-R	Emb-F
BiLSTM-CRF	0.8810	0.8628	0.8718	0.8928	0.8897	0.8913	0.4343	0.0960	0.1567
Flair-BiLSTM-CRF	0.8995	0.8843	0.8917	0.9294	0.8885	0.9084	0.8556	0.1360	0.2197
赛特	0.8999	0.8998	0.8998	0.9272	0.9021	0.9144	0.8489	0.1920	0.2947

和“Emb”（嵌入因果关系），代表因果关系的语义角色
事件，在我们的任务中起着至关重要的作用。为了进行误差分析，我们提出了一个
标签“C”（包括“BC”和“IC”）、“E”（包括“BE”和“IE”)和“Emb”（包括“B-
Emb”和“I-Emb”)的混淆矩阵如图9所示。
我们可以看到，大多数错误都是“C”、“E”和“O”之间的混淆。
这种混乱可能是由于注释数据不足的问题引起的。
与其他基线相比[5]，我们的模型 SCITE 可以更好地识别“C”，
“E”和“Emb”。

此外，我们将 SCITE 模型的标记性能与
基线。比较结果总结在表4中。首先，我们观察到
我们的模型在标签“C”（包括“BC”和“IC”）、“E”中获得第一
（包括“BE”和“IE”)和“Emb”（包括“B-Emb”和“I-Emb”）
F1 分数术语。其次，我们还注意到 F1 分数大约为
0.9，但标签“Emb”除外，因为其频率较低（仅 110 个实例）
训练集。特别是从图9的混淆矩阵可以看出
测试集中的大多数“Emb”标签被错误识别为“C”或“E”，这
导致“Emb”的召回率较低。

4.2.消融分析

研究 SCITE (Flair+CCNN-
BiLSTM-MHSA-CRF），我们还报告了 Ta- 中的消融实验结果
ble 5. 所有部分都对 SCITE 模型的性能做出积极贡献。

5为了展示方便，我们只展示了SICFI、Flair-BiLSTM-CRF的结果
（基线的优越者)和 BiLSTM-CRF（经典序列标记模型）。

表 5:我们提出的 SCITE 模型的消融分析。 “全部”表示完整的 SCITE 模型,即Flair+CCNN-BiLSTM-MHSA-CRF模型，“-”表示去掉来自 SCITE 的组件。

模型	环境	F
赛特	全部	0.8455
Flair-BiLSTM-MHSA-CRF	-CCNN	0.8438
Flair-BiLSTM-CRF	-CCNN -MHSA	0.8382
BiLSTM-MHSA-CRF	-天赋-CCNN	0.8137
BiLSTM-CRF	-天赋-CCNN-MHSA 0.7884	

具体来说,我们发现转移的 Flair 嵌入提供了最多的巨大的进步。这验证了我们的假设,即缺乏数据现有语料库中包含因果三元组会影响性能因果关系提取中基于神经的模型的研究。令人印象深刻的是,与没有 Flair 嵌入的 SCITE（SCITE-Flair）,转移的 Flair 嵌入该案例中 dings 的 F1 分数提高了 33.28%极度标注的数据不足（训练数据的 10%）,如图所示图 10. 在转移的上下文表示的帮助下,我们不仅可以从文本中学习更多的语义和句法信息还可以捕获上下文中的单词含义,以解决多义性和上下文问题词的依赖性。

此外,我们还发现多头自注意力（MHSA)机制无性主义可以进一步提高绩效,尤其是在没有天赋的情况下嵌入;原因在 4.3 节中讨论。最后,我们发现，特定于任务的角色特征也会影响模型的性能比较有角色和没有角色的模型时略有增加从 CCNN 中学习到的表示。

4.3.多头自注意力分析

与其他序列标记模型不同,SCITE 使用多头自注意力机制来学习因果关系。为了进一步分析 MHSA 的效果,我们计算并可视化 F1-

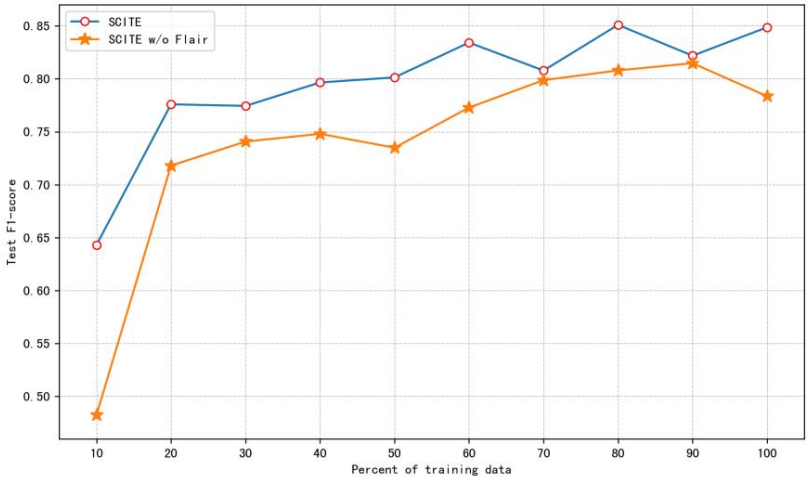


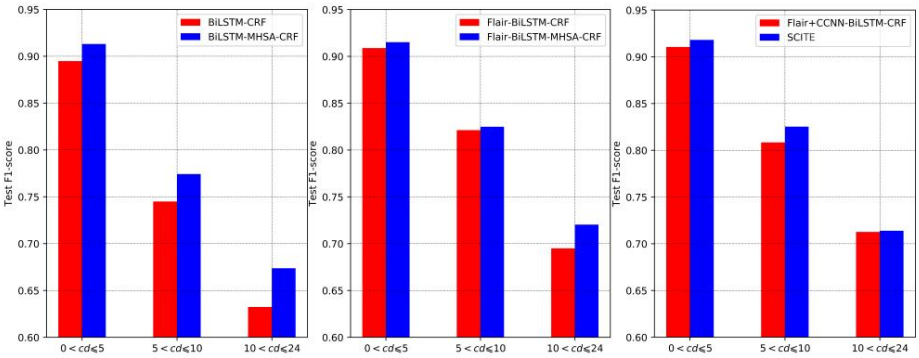
图 10:测试集上的 F1 分数（以训练数据集的大小表示）。

以因果关系距离（原因和结果之间的距离)表示的分数
三组模型：

- 第 1 组:BiLSTM-CRF 和 BiLSTM-MHSA-CRF；
- 第 2 组:Flair-BiLSTM-CRF 和 Flair-BiLSTM-MHSA-CRF；
- 第 3 组:Flair+CCNN-BiLSTM-CRF 和 SCITE（Flair+CCNN-BiLSTM-MHSA-CRF)

如图11所示,我们发现F1分数随着增加而减少
所有三组中的因果关系距离。这验证了我们的假设
因果关系之间的长期依赖导致了因果关系的困难
萃取。此外,我们还看到具有 MHSA 的模型的性能
在任意因果关系距离上优于没有 MHSA 的模型，
这表明 MHSA 机制在有效增强因果关联方面发挥着至关重要的作用。特别是,MHSA 显着
提高了因果距离方面的性能

与其他因果关系距离较短的情况相比,小于10,如图所示
图11a和图11b。



(a) 第 1 组 (b) 第 2 组 (c) 第 3 组

图 11:以因果关系距离 (cd) 表示的 F1 分数比较。我们划分
测试中的因果三元组根据其 cd 分为三部分:0 < cd ⩽ 5, 5 < cd ⩽ 10,
并且 10 < cd ⩽ 24（测试集中最大 cd 为 24）,比例约为 2:2:1。

4.4.案例分析

在表6中,我们列出了两个有代表性的例子来展示其优点和
我们提出的模型的缺点。对于每种情况,我们都会显示输入句子
以及第一行和第二行句子中包含的因果三元组。这
其余行显示不同模型提取的因果三元组 [5]。

第 1 句是简单因果关系的情况（参见第 2.2.1 节）,其中五个
因果三元组正在等待模型提取。我们观察到,无论是
SCITE模型或其他两个基线模型都无法获得所有因果三元组
正确。模型似乎很难了解“肤浅或
地下水”是一个完整的语义单位,短语“以及”
可能在连接两个因果成分方面发挥关键作用。原因可能是
基于我们的因果关系标记方案的序列标记模型需要
稍微多一点的训练数据来学习这些因果关系表达模式。

第 2 句是复杂因果关系的情况（参见第 2.2.2 节）,其中
是一种内在的因果关系,因此给解决问题带来了困难和模糊性
模型的因果关系学习。在此示例中,只有 SCITE 可以捕获所有
因果关系,从而精确提取所有三个
与其他模型相比,因果三元组。

表6:因果关系提取结果,其中“C”代表“原因”,“E”代表“影响”。我们用斜体显示正确的结果,并用粗体突出显示不正确的结果。

句子1	核实了[泥石流]、 [地震]、 [沉降]、 [地表或地下水]造成的[损害] ,以及[膨胀粘土]。 {[泥石流], CE, [损害]}, {[地震], CE, [损害]}, {[沉降], CE, [损害]}, {[地表或地下水], CE, [损害]}, {[膨胀粘土], CE, [损害]} {[泥石流], CE, [损害]}, {[震颤], CE, [损害]}, {[沉降], CE, [损害]}, {[地表], CE, [损害]}, {[地下水], CE, [损害]}
真正的三胞胎	
赛特	
Flair-BiLSTM-CRF	{[泥石流], CE, [损害]}, {[地震], CE, [损害]}, {[沉降], CE, [损害]}, {[地下水], CE, [损害]]} 没有任何
BiLSTM-CRF	{[泥石流], CE, [损害]}, {[地震], CE, [损害]}, {[沉降], CE, [损害]}, {[地下水], CE, [损害]]} 没有任何
句子2	今年的诺贝尔生理学或医学奖获得者 做出了非凡且意想不到的发现 胃[炎症]以及胃或十二指肠[溃疡]是由[幽门螺杆菌]引起的胃[感染]的结果。
真正的三胞胎	{[感染], CE, [炎症]}, {[感染], CE, [溃疡]}, {[幽门螺杆菌], CE, [感染]}
赛特	{[感染], CE, [炎症]}, {[感染], CE, [溃疡]}, {[幽门螺杆菌], CE, [感染]} {[幽门螺杆菌], CE , [炎症]}, {[幽门螺杆菌], CE, [溃疡]}, {[幽门螺杆菌], CE, [感染]}
Flair-BiLSTM-CRF	
BiLSTM-CRF	{[幽门螺杆菌],CE,[非凡且意想不到的发现]}, 没有任何, {[幽门螺杆菌],CE,[感染]}

5. 相关作品

在本节中,我们简要介绍提出的因果关系提取技术其他研究人员提出的方法,分为三类:1)采用的方法仅模式匹配,2)基于模式和组合的技术机器学习,3)基于深度学习技术的方法。

5.1.基于模式的方法

基于模式的方法通过使用se-模式匹配来提取因果关系曼蒂克特征、词汇句法特征和自构建约束。为了例如,Khoo 等人。 [1]从《华尔街日报》中提取因果知识-最后使用语言线索和模式匹配。在医疗领域摘要,Khoo 等人。 [11]使用图形模式来提取因果知识来自医学数据库。 Girju 和 Moldovan [15] 提取了因果关系将句法模式“NP1 因果动词 NP2”与使动动词组合,然后利用语义约束将候选者分类为因果或非因果。Ittoo和Bouma[16]提出了一种基于part-的因果对提取方法言语、句法分析和因果关系模板。在他们的工作中,因果关系首先使用维基百科上的因果句子提取模板,然后他们使用这些模板来提取其他句子中的因果关系。

这些仅依赖于模式匹配规则的方法通常具有跨领域适用性差,可能需要广泛的领域知识解决特定领域的问题,以及制定消耗的规则大量的时间和精力。

5.2.基于模式与机器学习相结合的方法

基于模式和机器学习技术相结合的方法niques主要以管道方式处理这个任务。他们首先提取候选人根据温度可能具有因果关系的短语（或实体、事件)对板块或一些线索词,然后根据候选因果对进行分类以一些统计特征或语义特征和语法特征来过滤ter非因果对。 Girju [5] 使用基于因果关系触发词的约束

提取英文文本中的因果关系并使用C4.5决策树
进行分类。索尔金特等人。 [17]使用预定义的模板来提取
候选因果对,然后使用贝叶斯分类器和拉普拉斯平滑
过滤非因果对。赵等人。 [18]提出了一个称为“因果关系”的新特征
连接词”通过计算句法依存结构的相似度
的句子。他们运行部分解析器来首先提取候选名词短语
然后使用受限隐藏朴素对候选因果对进行分类
贝叶斯学习算法结合其他特征,但他们的方法
无法区分原因和结果。罗等人。 [19] 提取原因-
使用因果线索从大规模网络文本语料库中影响术语,然后使用
一种新的基于统计指标的逐点互信息（PMI)来衡量
任意两段短文本之间的因果强度。

上述方法将因果关系提取分为两个子任务:候选任务
因果对提取和关系分类（过滤非因果对）。这
候选因果对提取的结果可能会影响关系的性能
化分类并产生级联错误。这些方法往往需要
在特征工程中投入大量人力和时间,严重依赖
手动选择文本特征,并且手动选择的特征是相关的
过于简单,无法捕获上下文的深入语义信息。

5.3.基于深度学习技术的方法

由于深度神经网络强大的表示学习能力
能够有效捕捉隐含的、模糊的因果关系的作品,
采用深度学习技术进行因果关系提取已成为流行趋势
近年来成为研究人员的选择。德席尔瓦等人。 [21]使用CNN对文本中的因果关系进行
分类。克鲁恩格莱等人。 [22]使用多列CNN
从嘈杂的文本中提取背景知识来对这些共同点进行分类
将因果关系理解为“吸烟”→“死亡或肺癌”。类似地,Li 和 Mao [24] 提出了一种结合先
验知识的面向知识的 CNN
从词汇知识库中进行因果关系分类。马丁内斯-卡马拉
等人。 [23]提出了一种基于 LSTM 的模型,仅使用词嵌入来进行

因果关系分类的任务。除了对因果关系进行分类之外
常识推理的立场,达斯古普塔等人。 [25] 和杜尼茨等人。 [26]
还从语言学角度识别了文本中因果关系的语言表达
通过基于 LSTM 的深度模型的观点。

我们提出的方法与上述方法的主要区别
基于深度学习技术可以概括如下:

- 我们的方法旨在自动提取此类常识性因果关系
  三元组如文本中的c（图1）,不仅是为了对因果关系进行分类或
  识别因果关系的语言表达。

- 我们的方法可以轻松处理多个因果三元组和嵌入式
  同一个句子中的因果关系（第 2.1 节和第 2.2 节）,但没有
  将句子分成仅包含一个实例的子句子
  因果关系,从而产生级联错误,如 Dasgupta 等人。 [25]。

六,结论

在本文中,我们将因果关系提取表述为序列标记问题
lem 并为因果关系提供基于 BiLSTM-CRF 的自注意力解决方案
萃取。特别是,我们建议 SCITE 来提取自然中的因果关系
基于我们的因果标记方案的语言文本。为了缓解问题
由于数据不足,我们将从大型数据集训练的 Flair 嵌入转移到
语料库到我们的任务中。此外,我们还介绍了多头自注意力
学习因果关系的机制。实验性的
结果证明了我们提出的方法的有效性。但是,那
SCITE 的性能仍然在一定程度上受到资源不足的限制。
高质量的注释数据（第 4.4 节）。

在今后的工作中,我们将尝试通过以下方式解决这个问题:

1. 基于现有数据集,开发多个来源的带注释数据集
   以及我们的因果关系标记方案。

2.将我们的方法与远程监督[55]和强化学习相结合-
荷兰国际集团[56]获得更好的性能,而无需建立一个高
用于因果关系提取的质量注释语料库。

7. 致谢

该研究得到国家自然科学基金委的部分支持
中国日期（编号:U1711263）。

参考

参考

[1] CSG Khoo、J. Kornfilt、RN Oddy、SH Myaeng、自动
从报纸文本中提取因果信息
基于知识的推理、文学和语言计算 13 (4)
（1998)177-186， doi:10.1093/llc/13.4.177。

[2] CSG Khoo、SH Myaeng、RN Oddy,使用因果关系
提高信息检索精度、信息处理与文本
管理 37 (1) (2001) 119–145， doi:10.1016/s0306-4573(00)00022-4。

[3] C. Silverstein、S. Brin、R. Motwani、JD Ullman,挖掘因果结构、数据挖掘和知识发
现的可扩展技术 4 (2/3)
(2000) 163–192， doi:10.1023/A:1009891813863。

[4] K. Radinsky、S. Davidovich、S. Markovitch,学习新闻因果关系
事件预测,见:第 21 届国际会议论文集
万维网 - WWW 12,ACM Press,909–918， doi:10.1145/2187836。
2187958, 2012。

[5] R. Girju,自动检测问答的因果关系，
见:ACL 2003 年多语言摘要研讨会论文集
和问答,计算语言学协会,76-83，
doi:10.3115/1119312.1119322,2003 。

[6]D.-S。张,K.-S。 Choi,增量提示短语学习和引导-
使用提示短语和词对概率进行因果关系提取的 ping 方法
能力,信息处理与管理 42 (3) (2006) 662–678，
doi:10.1016/j.ipm.2005.04.004。

[7] A. Sobrino、C. Puente、J. Olivas,从因果机械中提取答案
医学文献中的 anisms,Neurocomputing 135 (2014) 53–60， doi：
10.1016/j.neucom.2013.05.056。

[8] M. Riaz,R. Girju,因果关系的另一种视角:发现场景 -
无监督的特定应急关系,见:2010 IEEE
第四届语义计算国际会议,IEEE,361–368，
doi:10.1109/icsc.2010.19,2010 。

[9] C. Hashimoto、K. Torisawa、J. Kloetzer、M. Sano、I. Varga、J.-H。哦,Y.Ki-
dawara,走向未来场景生成:提取事件因果关系
利用语义关系、上下文和关联特征,见:Pro-
计算协会第52届年会会议记录
语言学（第一卷:长论文）,计算语言学协会,987–997， doi:10.3115/v1/p14-1093,2014 。

[10] EJM Ackerman,提取新闻主题的因果网络,载于:On
转向有意义的互联网系统:OTM 2012 研讨会,施普林格
柏林海德堡,33–42， doi:10.1007/978-3-642-33618-8 5,2012。

[11] CSG Khoo、S. Chan、Y. Niu,从医学中提取因果知识
使用图形模式的 ical 数据库,见:Proceedings of the 38th Annual
计算语言学协会会议 - ACL 00,Associa-
计算语言学,336–343， doi:10.3115/1075218.1075261，
2000年。

[12] 赵三、蒋明、刘明、秦本、刘涛,因果三元组:走向伪
从医学文本中发现因果关系并生成假设

数据来源:2018 年 ACM 国际生物化学会议论文集
格式学、计算生物学和健康信息学 - BCB 18、ACM
出版社,184–193， doi:10.1145/3233547.3233555,2018 。

[13] Y.丁,J.唐,F.郭,药物副作用关联的识别
具有居中内核对齐的多信息集成,神经-
计算 325 (2019) 211–224， doi:10.1016/j.neucom.2018.10.028。

[14] EC Alem´an Carre´on、H. Nonaka、A. Hentona、H. Yamashiro,测量
电视商业广告单纯曝光效应对购买的影响
基于机器学习预测模型的行为,Information Pro-
终止与管理 56 (4) (2019) 1339–1355, doi:10.1016/j.ipm.2019。
03.007。

[15] R. Girju,DI Moldovan,因果关系文本挖掘,见:Proceed-
第十五届国际佛罗里达人工智能研究会
学会会议,2002 年 5 月 14-16 日,美国佛罗里达州彭萨科拉海滩,360–
364,2002 年。

[16] A. Ittoo,G. Bouma,提取显性和隐性因果关系
来自稀疏的特定领域文本,位于:自然语言处理
和信息系统,施普林格柏林海德堡,52–63， doi:10.1007/
978-3-642-22327-3 6, 2011。

[17] A. Sorgente、G. Vettigli、F. Mele,因果关系自动提取
自然语言文本中的关系,见:第七届国际会议论文集
信息过滤和检索国际研讨会与
第十三届意大利人工智能协会会议 (AI*IA
2013）,意大利都灵,2013年12月6日。,2013年37-48日。

[18] 赵S.,刘T.,赵S.,Y.陈,J.-Y。聂, 基于连接词分析的事件因果关系提取, Neurocomputing 173
(2016) 1943–1950, doi:
10.1016/j.neucom.2015.09.066。

[19] Z. Luo,Y. Sha,KQ Zhu,S. Hwang,Z. Wang,常识因果关系
短文本之间的发音,见:知识表示原理
和推理:第十五届国际会议论文集,KR
2016,南非开普敦,2016 年 4 月 25-29 日。,421-431,2016。

[20]郑S.郑,王凤,鲍华,郝Y.周,P.周,徐B.,联合提取
基于新颖标签方案的实体和关系,见:Proceedings
计算语言协会第 55 届年会 -
抽动症（第一卷:长论文）,计算语言学协会,
1227–1236， doi:10.18653/v1/p17-1113,2017 。

[21] TN de Silva,X. 志博,Z. Rui,M. Kezhi,因果关系识别-
使用卷积神经网络和基于知识的特征,
世界科学、工程与技术学院,国际
计算机、电气、自动化、控制与信息学报
工程 11 (6) (2017) 697–702。

[22] C. Kruengkrai、K. Torisawa、C. Hashimoto、J. Kloetzer、J. Oh、M. Tanaka,
利用多种背景知识提高事件因果关系识别
使用多列卷积神经网络的边缘源,位于:Pro-
第三十一届 AAAI 人工智能大会论文集,
2017年2月4-9日,美国加利福尼亚州旧金山,3466-3473,2017。

[23] E. Mart´ nez-C´amara、V. Shwartz、I. Gurevych、I. Dagan,神经 Disam-
基于上下文的因果词汇标记的歧化,见:IWCS 2017 -
第十二届国际计算语义会议 - 短论文,
法国蒙彼利埃,2017 年 9 月 19 日至 22 日,2017 年。

[24] P. Li,K. Mao,面向知识的因果卷积神经网络
自然语言文本中的关系提取,专家系统与应用 115 (2019) 512–523， doi:10.1016/
j.eswa.2018.08.009。

[25] T. Dasgupta、R. Saha、L. Dey、A. Naskar,因果关系自动提取
使用语言学深度神经网络从文本中获取关系,

见:第 19 届 SIGdial 话语和会议年度会议记录
对话,澳大利亚墨尔本,2018年7月12-14日,306-316,2018。

[26] J. Dunietz、JG Carbonell、LS Levin、DeepCx:基于转换的 ap-
使用复杂的构造触发器进行浅层语义解析,
于:2018年自然经验方法会议论文集
语言处理,比利时布鲁塞尔,2018年10月31日至11月4日,
1691–1701, 2018。

[27] Z. Huang,W. Xu,K. Yu,序列的双向 LSTM-CRF 模型
标记,CoRR abs/1508.01991。

[28] I. Hendrickx、SN Kim、Z. Kozareva、P. Nakov、DO S´eaghdha、S. Pad´o、
M. Pennacchiotti、L. Romano、S. Szpakowicz、SemEval-2010 任务 8:多
名词对之间语义关系的方式分类,在:
第五届语义评估国际研讨会论文集,
SemEval@ACL 2010,乌普萨拉大学,瑞典乌普萨拉,7 月 15 日至 16 日,
2010,33-38, doi:10.3115/1621969.1621986,2010 。

[29] T. O Gorman、K. Wright-Bettner、M. Palmer、Richer 事件描述:
将事件共指与时间、因果和桥接注释相结合
化,在:第二届计算新闻故事研讨会论文集-
线(CNS 2016),计算语言学协会,47-56, doi:
10.18653/v1/w16-5706,2016 。

[30] N. Mostafazadeh、A. Grealish、N. Chambers、J. Allen、L. Vanderwende,
CaterS:语义注释的因果关系和时间关系方案
事件结构,见:第四届事件研讨会论文集,As-
计算语言学协会,51–61, doi:10.18653/v1/w16-1007,
2016年。

[31] A. Akbik、D. Blythe、R. Vollgraf,Se 的上下文字符串嵌入
序列标签,见:第 27 届国际会议记录
计算语言学,COLING 2018,圣达菲,新墨西哥州,美国,
2018年8月20日至26日,2018年1638年至1649年。

[32] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、AN Gomez，
L. Kaiser、I. Polosukhin，《注意力就是你所需要的一切》，载于：神经学进展
信息处理系统 30：神经信息年会
2017 年信息处理系统大会，2017 年 12 月 4-9 日，加利福尼亚州长滩，
美国，6000–6010，2017 年。

[33] Z. Tan，M. Wang，J. Xie，Y. Chen，X. Shi，深度语义角色标注
自我关注，见：第三十二届 AAAI 会议记录
人工智能的影响，（AAAI-18），第30届人工智能创新应用（IAAI-18），以及第八届AAAI研讨会

人工智能教育进展 (EAAI-18)，新奥尔良，
美国路易斯安那州，2018 年 2 月 2-7 日，4929–4936，2018。

[34] Y. LeCun、B. Boser、JS Denker、D. Henderson、RE Howard、W. Hub-
吟游诗人，LD Jackel，反向传播应用于手写邮政编码
识别，神经计算 1 (4) (1989) 541–551， doi:10.1162/neco。
1989.1.4.541。

[35] X. Ma, E. Hovy，通过双向 LSTM 进行端到端序列标记 -
CNNs-CRF，见：协会第 54 届年会记录
计算语言学（第一卷：长论文），协会
计算语言学，1064–1074， doi:10.18653/v1/p16-1101，2016 。

[36] R. Collobert、J. Weston、L. Bottou、M. Karlen、K. Kavukcuoglu、PP
Kuksa，《自然语言处理（几乎）从头开始》，《自然语言处理杂志》
机器学习研究 12 (2011) 2493–2537。

[37] M. Peters、M. Neumann、M. Iyyer、M. Gardner、C. Clark、K. Lee、L. Zettle-
moyer，深度语境化的词表示，见：Proceedings of
2018年协会北美分会会议
计算语言学：人类语言技术，第 1 卷
（长论文），计算语言学协会，2227–2237， doi：
10.18653/v1/n18-1202，2018 。

[38] J. Devlin、M. Chang、K. Lee、K. Toutanova、BERT:预训练
用于语言理解的深度双向转换器,CoRR
绝对/1810.04805。

[39] C. Chelba、T. Mikolov、M. Schuster、Q. Ge、T. Brants、P. Koehn、T. Robin-
儿子,衡量统计领域进展的十亿字基准
量表建模,见:INTERSPEECH 2014,第 15 届年会
国际言语交流协会,新加坡,九月
14-18, 2014, 2635–2639, 2014.

[40] A. Komninos、S. Manandhar,基于依赖的句子嵌入
分类任务,见:2016 年北方会议记录
计算语言学协会美国分会:人类语言技术,计算语言学协会,

1490–1500， doi:10.18653/v1/n16-1175,2016 。

[41] S. Hochreiter,J. Schmidhuber,长短期记忆,神经计算机
9 (8) (1997) 1735–1780, doi:10.1162/neco.1997.9.8.1735。

[42] Y. Bengio、P. Simard、P. Frasconi,通过梯度下降学习长期依赖关系很困难,IEEE
Transactions on Neural Networks 5 (2)
（1994)157-166， doi:10.1109/72.279181。

[43] J. Cheng、L. Dong、M. Lapata,长短期记忆网络
机器阅读,见:2016 年实证会议论文集
自然语言处理方法,计算协会
语言学,551–561， doi:10.18653/v1/d16-1053,2016 。

[44] Z. Lin,M. Feng,CN dos Santos,M. Yu,B. Xiang,B. Zhou,Y. Bengio，
结构化自注意力句子嵌入,CoRR abs/1703.03130。

[45] JD Lafferty、A. McCallum、FCN Pereira,条件随机场：
用于分割和标记序列数据的概率模型,位于:Pro-
第十八届国际机器学习会议论文集

(ICML 2001),威廉姆斯学院,美国马萨诸塞州威廉斯敦,6 月 28 日 - 7 月 1,2001 年,282–289,2001 年。

[46] AJ Viterbi,卷积码和渐近线的误差界限
最佳解码算法,IEEE Trans。信息论 13 (2) (1967) 260–269。

[47] N. Reimers、I. Gurevych,报告分数分布带来不同 -
ence:用于序列标记的 LSTM 网络的性能研究,位于:
2017年自然兰实证方法会议论文集
语言处理,计算语言学协会,338–348， doi:
10.18653/v1/d17-1035,2017 。

[48] Y. Gal, Z. Ghahramani,Dropout 的理论应用
在循环神经网络中,在:神经信息专业进展
cessing Systems 29:神经信息处理年会
Systems 2016,2016 年 12 月 5-10 日,西班牙巴塞罗那,1019-1027,2016。

[49] R. Pascanu、T. Mikolov、Y. Bengio,论循环训练的难度
神经网络,见:第 30 届国际会议论文集
机器学习,ICML 2013,美国佐治亚州亚特兰大,2013 年 6 月 16-21 日,1310–1318,2013。

[50] T. Dozat,将 Nesterov Momentum 融入 Adam,载于:Proceedings
第四届学习表征国际会议,研讨会
轨道,2016。

[51] E. Strubell、P. Verga、D. Belanger、A. McCallum,快速准确的实体
使用迭代扩张卷积进行识别,见:Proceedings of the
2017年自然语言处理经验方法会议，
计算语言学协会,2670–2680， doi:10.18653/v1/
d17-1283,2017年。

[52] 王平,钱勇,宋福康,何丽,赵红,词性标注

具有双向长短期记忆循环神经网络，
CoRR 绝对/1510.06168。

[53] G. Lample、M. Ballesteros、S. Subramanian、K. Kawakami、C. Dyer、Neu-
命名实体识别的真实架构，见：Proceedings of the
2016年协会北美分会会议
计算语言学：人类语言技术，协会
计算语言学，260–270， doi:10.18653/v1/n16-1030，2016 。

[54]L。 Borchmann、A. Gretkowski、F. Grali´nski，接近嵌套命名 en-
使用并行 LSTM-CRF 进行实体识别，见：Proceedings of the PolEval
2018 年研讨会，63-73，2018。

[55] M. Mintz、S. Bills、R. Snow、D. Jurafsky，关系的远程监督
无标签数据提取，见：联席会议记录
ACL第47届年会暨第四届国际联合会议
AFNLP 自然语言处理会议：第 2 卷 -
ACL-IJCNLP 09，计算语言学协会，1003–1011，
doi：10.3115/1690219.1690287，2009 。

[56] RS Sutton、AG Barto，强化学习：简介，IEEE
跨。神经网络 9 (5) (1998) 1054–1054, doi:10.1109/TNN.1998。
712192。