基于SoftLexicon和注意力机制的中文因果关系抽取

崔仕林[1,2,3],闫蓉[1,2,3*]
[1]内蒙古大学计算机学院/内蒙古, 呼和浩特, 010021
[2]蒙古文智能信息处理技术国家地方联合工程研究中心/内蒙古, 呼和浩特, 010021
[3]内蒙古自治区蒙古文信息处理技术重点实验室/内蒙古, 呼和浩特, 010021
1437869230@qq.com, csyanr@imu.edu.cn

摘要

针对现有中文因果关系抽取方法对因果事件边界难以识别和文本特征表示不充分的问题，提出了一种基于外部词汇信息和注意力机制的中文因果关系抽取模型BiLSTM-TWAM+CRF。该模型首次使用SoftLexicon方法引入外部词汇信息构建词集，解决了因果事件边界难以识别的问题。通过构建的双路关注模块TWAM(Two Way Attention Module)，实现了从局部和全局两个角度充分刻画文本特征。实验结果表明，与当前中文因果关系抽取模型相比较，本文方法表现出更优的抽取效果。

关键词： 因果关系抽取 ； 序列标注 ； 外部词汇信息 ； 注意力机制

Chinese Causality Extraction Based on SoftLexicon and Attention Mechanism

Shilin Cui[1,2,3],Rong Yan[1,2,3*]
[1]College of Computer Science, Inner Mongolia University, Hohhot, 010021, China
[2]National and Local Joint Engineering Research Center of
Intelligent Information Processing Technology for Mongolian, Hohhot, 010021, China
[3] Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot, 010021, China
1437869230@qq.com, csyanr@imu.edu.cn

Abstract

Existing Chinese causality extraction methods have to face the problems of identifying causal event boundaries and inadequate text features representation, this paper proposes a Chinese causality extraction model BiLSTM-TWAM+CRF based on external lexical information and attention mechanism for addressing above issues. It is the first time that we introduce external lexical information by using the SoftLexicon method to construct word set for solving causal event boundaries problem. We construct a Two Way Attention Module (TWAM) and try to represent the text features as much as possible from both the local and global views. Experimental results show that our proposed method has better causality extraction performance than the existing Chinese causality extraction methods.

Keywords: Causal Relation Extraction , Sequence Labeling , External Lexical Information , Attention Mechanism

1 引言

基金项目：国家自然科学基金(61866029)

随着海量数据的增长，如何快捷地从海量文本中寻找有用信息已经成为一项研究难题，因此信息抽取研究应运而生。因果关系抽取研究 (杨竣辉et al., 2016)作为其重要分支，多年来受到了学者们的广泛关注，在自然语言处理(Natural Language Processing,NLP)领域研究中占有重要地位。

因果关系抽取任务旨在从自然语言文本中自动抽取出文本中的因果关系。因果关系是文本中重要的一种语义关系，大量的存在于自然语言文本中。它常被用来描述‘原因’与‘结果’之间的前后联系 (冯冲et al., 2018)，在事件推理 (Radinsky et al., 2012)、未来场景生成 (Hashimoto et al., 2014)、问答 (Girju, 2003)和信息检索等任务中起着十分重要的作用。同时，根据文本中是否含有因果连接词，可以将因果关系分为显式因果关系和隐式因果关系。例如：①‘火灾导致两名儿童受伤。’中，‘导致’为因果连接词，‘火灾’和‘受伤’构成显式因果关系；②‘男子盗窃被抓获。’中，不含有因果连接词，‘ 盗窃’和‘被抓获’构成隐式因果关系。

目前，因果关系抽取任务的相关研究，主要从三个角度展开：①使用文本分类，判断句子中是否具有因果关系；②给定候选因果对，判断句子中是否包含因果关系；③通过序列标注方法，对句子抽取因果关系并确定因果关系的方向。这些研究方法虽然都能够对因果关系实现有效地抽取，但是依然存在以下问题。首先，中文文本与英文文本最大的区别在于没有明显的边界标示符，所以对中文文本处理需要先确定中文词汇的边界，即需要进行中文分词。同时，中文词语的一词多义等问题会导致词语边界模糊，很难通过分词工具得出准确的词语边界。其次，由于显式因果关系含有明显的因果连接词，大多数的模型只对显式因果关系做到了抽取，而隐式因果关系由于缺乏明显的词汇特征，使得大多数的模型对其抽取效果较差。最后，针对中文文本，单纯基于字符向量的方法无法利用中文词汇信息，致使现有的单纯基于字符向量的方法对文本的特征表示不充分。

最近的相关研究表明，引入外部词汇信息 (Zhang and Yang, 2018)和注意力机制 (Vaswani et al., 2017)，能够在一定程度上解决中文词汇边界难以识别和文本特征表示不充分的问题。本文延续这一思路，使用基于序列标注的方法，抽取中文文本中的因果关系。具体地，在字符向量的基础上，利用SoftLexicon方法 (Ma et al., 2020)引入外部词汇信息，通过构建词集的方式，将词典信息融合到字符表示层中，增强了字符的语义表达，从而解决中文词汇边界难以识别的问题。进一步地，本文结合注意力机制构建了双路关注模块TWAM(Two Way Attention Module)，该模块融合了通道注意力(Channel Attention) (Hu et al., 2020)和缩放点积注意力(Scaled Dot-Product Attention) (Vaswani et al., 2017)，能够从局部和全局角度捕获句子的语义特征，提取出深层次的语义信息，进而增强文本的特征表示。同时结合BiLSTM与CRF，本文构建了BiLSTM-TWAM+CRF模型，该模型不仅引入了外部词汇信息以解决无法利用词汇信息的问题，而且利用构建的TWAM，从局部和全局两个角度捕获了更多的文本特征，能够更加有效地抽取中文文本中的因果关系。

2 相关工作

2.1 因果关系抽取

现有抽取因果关系的方法主要分为三种：基于规则的方法 (Garcia, 1997; Ittoo and Bouma, 2011)、基于统计的方法 (Zhao et al., 2016; Nauta et al., 2019)和基于深度学习的方法 (Zhang et al., 2015; Jin et al., 2020)。基于规则的方法使用模式匹配抽取因果关系，根据文本结构特征，人为制定规则，虽然准确度高，但泛化能力和可移植性较差。基于统计的方法虽然通过抽象得到的文本特征克服了依赖领域规则的问题，但它需要复杂的特征工程，很大程度上依赖于标注语料的质量，并且人工特征工程会带来额外的噪声，从而影响抽取精度。

近些年来，由于基于深度学习的方法能够从自然语言文本中自动学习文本特征，有效地解决了跨领域抽取及人工干预等问题，涌现了大量利用深度学习技术来抽取文本中因果关系的研究。文献 (Zeng et al., 2014)使用卷积神经网络CNN(Convolutional Neural Networks)对词汇向量和位置向量提取特征信息，提高了关系抽取的性能。但是，由于CNN不适合处理长距离的上下文语义特征，进一步地，文献 (Zhang et al., 2015)提出使用双向长短期记忆网络BiLSTM(Bidirectional Long Short-term Memory Networks)提取词汇特征，利用BiLSTM网络中的LSTM(Long Short-Term Memory,长短期记忆网络)单元实现了对文本关系的抽取。虽然BiLSTM可以对句子的长距离依赖关系进行建模，但是单纯的BiLSTM提取到的语义特征并不充分。文献 (Zeng et al., 2016) 提出了卷积双向LSTM模型，该模型利用双向LSTM提取句子

级别特征，利用CNN抽取单词在句子中的局部上下文特征，既增强了句子中的语义信息表示，又避免了人工特征导致的噪声问题，实现了对因果关系的抽取。预训练模型BERT (Jacob et al., 2018)出现后，文献 (Gao et al., 2021)提出了一种基于BERT预训练模型的因果关系联合提取模型，该模型使用BERT增强句子的语义表示，利用迭代扩展卷积增强事件的因果关系，从而实现对事件内部因果关系的抽取。

另外，还有一些因果关系抽取研究是基于序列标注方法展开的。文献 (姜博et al., 2021b)基于BiLSTM+CRF (Huang et al., 2015)和BERT预训练模型提出了BERT+BiLSTM+CRF方法，使用BERT预训练模型增强句子中字符的特征信息，可以有效地抽取文本中的因果关系。文献 (郑巧夺et al., 2021c)联合了卷积神经网络与双向门控循环单元BiGRU(Bidirectional Gated Recurrent Unit)，使用两次序列标注任务实现了因果关系的抽取，并且结合BERT预训练模型，增强了文本特征的表达能力，提升了模型对语义特征的提取能力。文献 (Li et al., 2021)提出了SCIFI因果关系抽取器，将上下文字嵌入应用到因果关系抽取任务中，并且使用多头注意力机制学习因果关系词之间的依赖关系，实现了对因果关系的直接提取。与传统的因果关系抽取方法相比，基于深度学习方法的因果关系抽取模型效果有明显提升，而且采用序列标注的方法，能够实现真正的因果关系抽取。

2.2 SoftLexicon

近些年来，许多基于序列标注方法对中文因果关系抽取的研究，通常采用基于字符的方式。但是由于中文文本预处理需要分词，使得基于字符的方式不可避免地会引入错误的分词信息，导致词汇的边界模糊和标注错误。针对这一问题，很多研究开始在序列标注任务中引入外部词汇信息，以增强基于字符的特征表示。文献 (Zhang and Yang, 2018) 提出了Lattice-LSTM模型，将字符与字符匹配到的词汇融合，用以引入外部词汇信息。但是该模型结构复杂、训练和推理速度慢，并且不具备可迁移性。为了解决可迁移性和计算复杂问题，文献 (Ma et al., 2020)提出了一种引入词汇信息的简易方法SoftLexicon，将字符与词典进行匹配得到与字符相应的匹配词，然后按照字符在词语中的位置，将匹配得到的词语分别放置在四个词集$\{B, M, E, S\}$中，尽可能保留了所有的词典匹配结果。其中，四个词集$\{B, M, E, S\}$分别表示该字符在词语中的开头、中间、结尾和单独构成一个字。如果在词典匹配后，词集为空，则以'None'填充该词集。为了能解决中文因果关系抽取方法对因果事件边界难以识别的问题，本文首次在中文因果关系抽取任务中引入外部词汇信息，利用SoftLexicon方法将字符信息与词汇信息融合，不仅利用了词汇的边界信息，还利用了词汇的语义信息，用以提升中文因果关系抽取的能力。

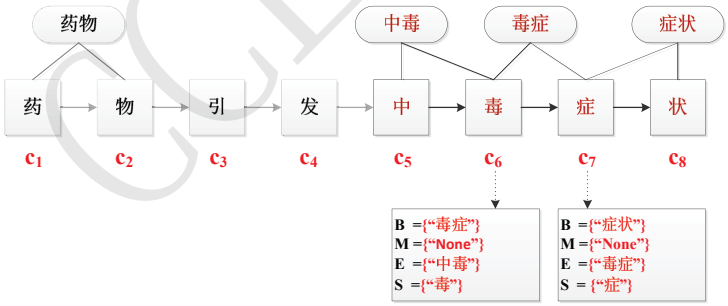


图 1: Softlexicon匹配示意图

如图 1所示，以'中毒症状'为例，字符'毒'与词典匹配可以得到三个词语'中毒'、'毒症'和'毒'。按照字符'毒'在这三个词语中的位置，于是，我们可以得到对应的四个词集，B={'毒症'}，M={'None'}，E={'中毒'}，S={'毒'}。

2.3 双向长短期记忆网络(BiLSTM)

循环神经网络RNN(Recurrent Neural Networks)自问世以来，由于其能够考虑上下文信息，被广泛用来处理时序性序列，但是该网络容易造成梯度消失和梯度爆炸。文献 (Hochreiter and Schmidhuber, 1997)在RNN基础上构建了长短期记忆神经网络(LSTM)，它既可以处理时序性序列，又可以缓解梯度消失问题。但LSTM只能根据前一个时刻预测下一个时刻，文

献 (Graves and Schmidhuber, 2005)在LSTM的基础上提出了双向长短期记忆网络(BiLSTM)，它由正向LSTM和反向LSTM组成，从两个方向建模句子的上下文信息，正向LSTM从前向后获取特征，后向LSTM从后向前获取特征。

2.4 注意力机制

近年来，许多研究在因果关系抽取任务中使用了注意力机制。文献 (Nauta et al., 2019)提出了一种基于注意力机制和卷积神经网络的因果关系抽取模型，通过注意力机制捕捉句子的上下文特征，完成了对时序性数据中因果关系的抽取。文献 (Jin et al., 2020)构建了级联结构神经网络CSNN，利用自注意力机制来挖掘句子内部的语义特征，实现了对文本中因果关系的抽取。文献 (Gao et al., 2021)建立了领域知识融合模型，利用注意力机制对因果知识建模，以捕获事件内部的语义特征，挖掘特征之间的内部因果关系。

3 BiLSTM-TWAM+CRF模型

本文提出的BiLSTM-TWAM+CRF模型结构主要由三部分组成，分别是嵌入层、BiLSTM-TWAM层和标签预测层，整体结构如图 2所示。首先，将输入的句子转换成基于字符的向量和基于SoftLexicon方法的向量，并融合两个向量。其次，将嵌入层得到的向量输入至BiLSTM-TWAM层中，使用BiLSTM 提取文本的长距离语义特征，并构建双路关注模块TWAM(Two Way Attention Module)学习字符之间的依赖关系，然后使用残差结构将BiLSTM和TWAM的输出融合，得到最终的语义特征。最后，将最终的语义特征输入到标签预测层中，计算序列的最优解，输出最优结果。接下来的部分将详细说明各个部分。

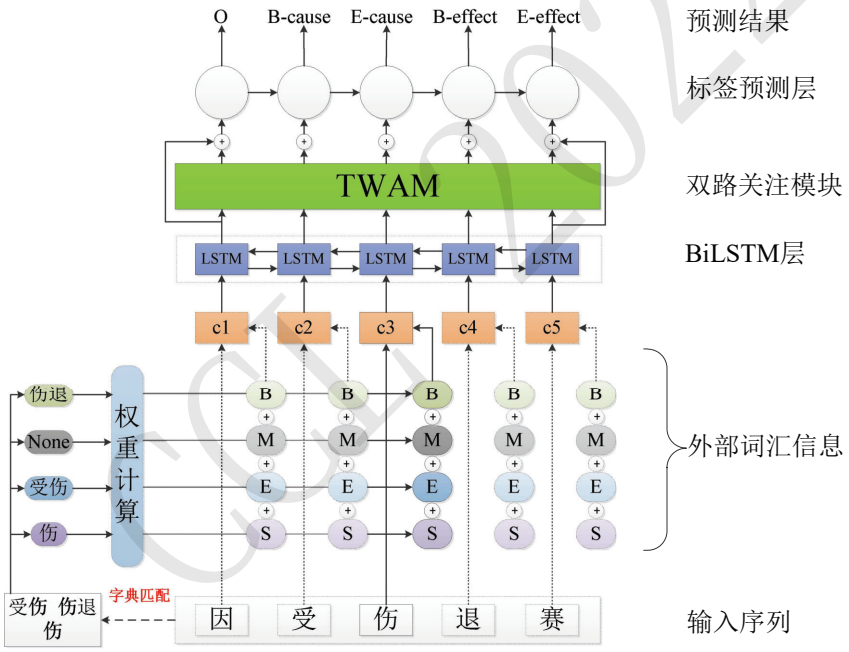


图 2: BiLSTM-TWAM+CRF模型结构

3.1 BiLSTM-TWAM层

本文提出的双路关注模块(TWAM)结构如图 3所示。BiLSTM-TWAM层由BiLSTM和TWAM组成，首先使用BiLSTM获得句子的长距离语义特征，然后利用TWAM提取深层次的语义信息，最后采用残差结构融合BiLSTM和TWAM的输出。

TWAM采用通道注意力和缩放点积注意力在局部和全局两个角度学习字符之间的依赖关系，使之能够更加充分地刻画文本特征。其中，通道注意力是按照通道对映射特征进行建模，将各通道的空间信息特征作为各通道的表示，使用池化操作聚合空间信息，提取出句子的局部特征。缩放点积注意力是对全部的映射特征进行建模，捕获到特征内部的相关性以及长距离依赖，提取出句子的全局特征。

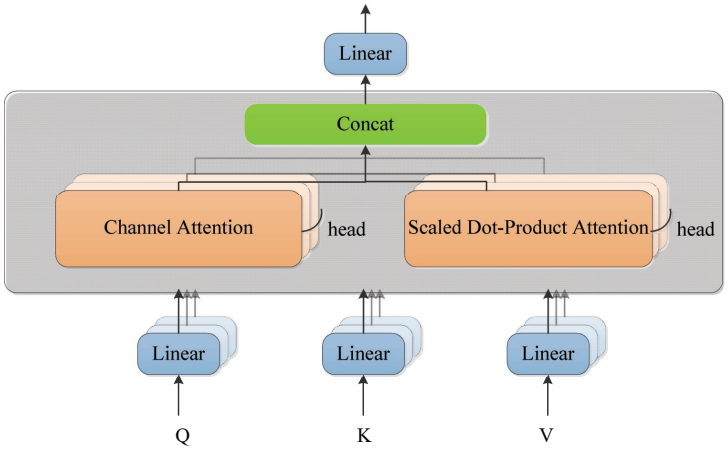


图 3: TWAM结构图

BiLSTM-TWAM层主要实现过程如下：

步骤一：将嵌入层得到的向量x^c输入BiLSTM中，对句子的长距离上下文语义特征进行提取得到语义特征表示H，然后将提取的特征表示H输入到TWAM 中。

步骤二：在TWAM得到来自BiLSTM的输出H后，建立查询矩阵Q、键矩阵K和值矩阵V，令Q=K=V=H。为获得更加丰富的语义信息，将Q矩阵、K矩阵和V矩阵分别映射至head个不同的子空间中，再输入至head个并行头中。

步骤三：每个并行头接收到不同子空间内的Q、K、V矩阵后，利用通道注意力和缩放点积注意力对不同子空间内的特征进行聚合，以达到关注不同子空间信息的目的。

首先，通道注意力使用平均池化和最大池化操作来聚合映射的空间信息特征，生成最大池化特征和平均池化特征，然后将两个特征送入多层感知器(MLP) 中，将MLP输出的特征进行融合，再经过sigmoid激活函数得到通道注意力结果，即文本的局部信息，计算如公式 1所示。

$$M_C(Q) = \sigma(MLP(AvgPool(Q) + MaxPool(Q))) \tag{1}$$

其中，σ为sigmoid函数，$AvgPool(Q)$和$MaxPool(Q)$表示平均池化特征和最大池化特征。

缩放点积注意力使用点积来计算Q矩阵与K矩阵的相似度，再除以$\sqrt{d_k}$(其中d_k为矩阵Q的维度)，并使用softmax函数计算权值，之后再乘以V矩阵得到缩放点积注意力结果，即文本的全局信息，计算如公式 2所示。

$$SDPA(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{2}$$

接着将每个并行头通过通道注意力得到的局部信息和缩放点积注意力得到的全局信息融合作为该并行头的结果，如公式 3所示。

$$\text{head}_\text{i} = (M_i(Q) \oplus SDPA(QW_i^Q, KW_i^K, VW_i^V)) \tag{3}$$

其中，W_i^Q、W_i^K和W_i^V分别为矩阵Q、K和V在第i个子空间内的权重矩阵，$\bigoplus$为拼接操作。

步骤四：将head个并行头得到的结果进行拼接后，再通过一个线性映射层得到TWAM的特征表示，计算如公式 4所示。

$$TWAM(Q, K, V) = Concat(head_1, \cdots, head_h)W^o \tag{4}$$

步骤五：最后本文使用残差结构将TWAM的特征表示与BiLSTM的特征表示融合，生成最终的语义特征表示。

3.2 嵌入层

对于有n个字符的输入序列s=$\{c_1, c_2, \cdots, c_n\}$，每个字符$c_i$通过嵌入可以得到该字符的字符向量$x_{\mathrm{i}}^c$，如公式 5 所示。

$$x_{\mathrm{i}}^c = e^c(c_i) \tag{5}$$

其中e^c表示字符嵌入查找表。

利用SoftLexicon方法引入外部词汇信息，将输入序列s的每个字符与词典匹配，得到序列s所有的匹配词。然后将匹配词按照$\{B, M, E, S\}$划分为四个词集，$B(c_i)$表示以c_i开始的词集，$M(c_i)$表示c_i位于词语中间的词集，$E(c_i)$ 表示以c_i结尾的词集，$S(c_i)$表示c_i单独构成的词集，四个词集的表示如公式 6 所示。

$$\begin{cases} B(c_i) = \{\mathrm{w}_{i,k}, \forall w_{i,k} \in L, i < k \leq n\}, \\ M(c_i) = \{\mathrm{w}_{j,k}, \forall w_{j,k} \in L, 1 \leq j < i < k \leq n\}, \\ E(c_i) = \{\mathrm{w}_{j,i}, \forall w_{j,i} \in L, 1 < j \leq i\}, \\ S(c_i) = \{c_i, \exists c_i \in L\}. \end{cases} \tag{6}$$

其中，L表示本文使用的外部词典。

获得每个字符的$\{B, M, E, S\}$四个词集后，需要将每个词集压缩成一个固定的向量，在压缩过程中，我们用词频作为权重，进行动态加权处理，通过计算得到词集S的词集向量，计算如公式 7所示。

$$\begin{cases} v^s(S) = \frac{4}{Z} \sum\limits_{w \in S} z(w) e^w(w) \\ Z = \sum\limits_{\mathrm{w} \in B \cup M \cup E \cup S} z(w) \end{cases} \tag{7}$$

其中，$z(w)$表示词典中词w出现的频率，Z表示词集中所有词出现的频率之和，$v^s(S)$表示压缩后的集合向量。

完成词集的向量化后，采用向量拼接的方式，将四个词集添加到每个字符的表示中，每个字符的最终表示如公式 8所示。

$$\begin{cases} e^s(B, M, E, S) = [v^s(B), v^s(M), v^s(E), v^s(S)], \\ x^c \leftarrow [x^c; e^s(B, M, E, S)] \end{cases} \tag{8}$$

其中，x^c为融合词汇信息后的向量表示。

3.3 标签预测层

标签预测层采用的是条件随机场模型(Conditional Random Field,CRF)(Lafferty et al., 2001)。CRF是一种判别式的无向图模型，通过研究标签之间的关系，获得全局最优的标签序列。假设序列y=$\{y_1, y_2, \cdots, y_n\}$是给定输入句子$x$=$\{x_1, x_2, \cdots, x_n\}$的标签序列，则该序列的CRF分数计算如公式 9所示。

$$score(x, y) = \sum_{i=1}^{n} p_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} \tag{9}$$

其中，p_{i,y_i}表示句子中第i个字符为y_i标签的预测概率，A_{y_{i-1}, y_i}表示标签y_{i-1}到标签y_i的转移概率。

本文使用Viterbi(Forney et al., 1973)算法输出具有最大$score(x, y)$的标签序列，当损失函数最小时，表示该模型已经收敛，损失函数如公式 10所示。

$$E = log \sum_{y \in \mathrm{Y}} \exp^{s(y)} - score(x, y) \tag{10}$$

其中，Y表示句子x可能对应的标签序列的集合。

第二十一届中国计算语言学大会论文集，第190页-第200页，南昌，中国，2022年10月14日至16日。

4 实验与结果分析

4.1 实验数据集

本文实验选取中文突发事件数据集(Chinese Emergency Corpus,CEC)[0]和百度中文事件抽取数据集(DUEE)(Xinyu et al., 2020)作为语料。CEC数据集由上海大学语义智能实验室构建，包括地震、火灾、交通事故、恐怖袭击和食物中毒五个类别，借助互联网收集CEC生语料，按照XML语言对语料进行标记。DUEE数据集是百度发布的中文事件抽取数据集，从百度资讯信息文本中收集语料，共有六十五个事件类型。

本文采用'BMESO'序列标注方法对数据集中文本含有的因果事件进行标注。其中，'B-'表示该字在事件的开头，'M-' 表示该字在事件的内部，'E-'表示该字在事件的结尾，'S-' 表示该字本身为一个事件，'-cause'表示为原因事件，'-effect'表示为结果事件，'O' 表示无关字符。由于CEC数据集使用XML语言进行标记，首先对数据集去除HTML标签进行了格式处理，提取文本数据并获取其*Participant*、*Time*、*Denoter*和*Location*标签作为因果事件标注的依据。接着，用人工标注的方法，对获取到的文本数据标注出原因事件、结果事件和其他，最终从CEC 数据集提取出了1,026 条样本数据。对于DUEE数据集，首先获取其*Text*标签作为文本数据，同时将其*Event-Type*、*Trigger*和*Class* 标签作为因果事件标注的依据，再仿照CEC数据集，使用人工标注的方法对文本数据标注，最终提取出4,800条样本数据。两个数据集完成人工标注工作后，按照7:1:2的数量比例将两个数据集划分为训练集、验证集与测试集。CEC 数据集与DUEE数据集详细信息如表 1所示。

	CEC	DUEE
因果事件对	844	2,805
原因事件	898	2,938
结果事件	1,223	3,348

表 1: CEC和DUEE数据集描述

4.2 实验参数设置

本文用word2vec (Tomas et al., 2013)训练得到字符向量和词向量，选用Adam (Kingma and Ba, 2015)作为优化器，参数设置如表 2所示。

参数	参数设置
字向量维度	50
词向量维度	50
学习率	0.005
迭代次数	50
批大小	16
Dropout	0.5
BiLSTM隐藏单元数	100
TWAM中head	4

表 2: 实验参数

4.3 评价指标

本文实验根据句子中抽取得到的因果事件对是否正确来判定模型的抽取性能，即对于一组因果关系，若原因事件与结果事件均抽取正确，则因果关系抽取正确，否则抽取错误。本文将抽取因果事件对的准确率P、召回率R和$F1$值作为评价指标。

[0]https://github.com/shijiebei2009/CEC-Corpus

4.4 基线

为验证本文模型的有效性，本文与近几年提出的因果关系抽取模型和序列标注模型进行了对比实验。

(1) CNN-BiGRU (苗佳et al., 2021a)：用于事件触发词抽取，利用CNN提取词汇特征，使用BiGRU提取句子特征。

(2) CSNN (Jin et al., 2020)：级联结构模型，结合CNN和具有自注意力机制的LSTM对因果关系进行抽取。

(3) BERT+CNN-BiGRU (郑巧夺et al., 2021c)：基于残差思想的双层因果关系抽取方法，使用BERT提取语义特征，使用双层CNN-BiGRU模型增强语义表征能力。

(4) BiLSTM+CRF (Huang et al., 2015)：经典的序列标注模型，由BiLSTM与CRF分类器构成。

(5) SoftLexicon+BiLSTM+CRF (Ma et al., 2020)：字词联合的序列标注模型，利用SoftLexicon方法引入了词汇信息，在序列标注任务中取得了好的效果。

4.5 结果分析

4.5.1 不同方法的对比

本文提出的BiLSTM-TWAM+CRF模型采用 4.2节的参数设置，对比模型使用的参数设置都参考其原论文中的描述，对比实验结果如表 3 所示。

	CEC			DUEE		
模型	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$(%)
CNN-BiGRU	74.65	61.44	67.37	70.39	57.03	62.99
CSNN	70.55	54.87	61.72	71.99	62.10	66.56
BiLSTM+CRF	69.32	63.70	66.28	65.26	52.51	58.08
SoftLexicon+BiLSTM+CRF	74.00	65.33	69.33	70.48	**71.68**	71.04
本文模型	**76.43**	**71.90**	**74.04**	**76.40**	71.63	**73.80**
BERT+CNN-BiGRU	73.35	72.38	72.78	72.98	65.67	69.01
BERT+本文模型	**80.27**	**76.42**	**78.39**	**77.70**	**72.23**	**74.86**

表 3: 对比实验结果

从表 3可以看出，本文提出的BiLSTM-TWAM+CRF模型在CEC数据集和DUEE数据集上都取得了较好的效果。其中，对比没有引入外部词汇信息的模型，序列标注模型SoftLexicon+BiLSTM+CRF在三个评价指标上均表现突出，表明了利用SoftLexicon引入外部词汇信息对中文因果关系抽取任务的有效性。我们分析这主要是因为引入外部词汇信息后，与单纯基于字符向量的模型相比，基于字词联合的模型不仅能够利用词汇的边界信息，还可以利用词汇的语义信息，从而有效避免了单纯基于字符向量的方法不能够准确确定词语边界和标注错误的问题。

同时也可以观察到，在CEC数据集上，本文模型的效果优于序列标注模型SoftLexicon+BiLSTM+CRF，表明TWAM能够提升模型的中文因果关系抽取能力。TWAM中的通道注意力和缩放点积注意力，能够在局部和全局两个角度提取特征信息，可以更加充分的刻画句子的语义特征。TWAM中残差结构的引入也使得BiLSTM-TWAM层既能获取文本的上下文信息，又能对特征进行更深层次的特征提取。从表 3可以看到，在DUEE数据集上准确率P 和$F1$值分别提高了6.08%和2.76%，召回率R仅降低了0.05%，表明本文提出的TWAM在中文因果关系抽取任务中的有效性。

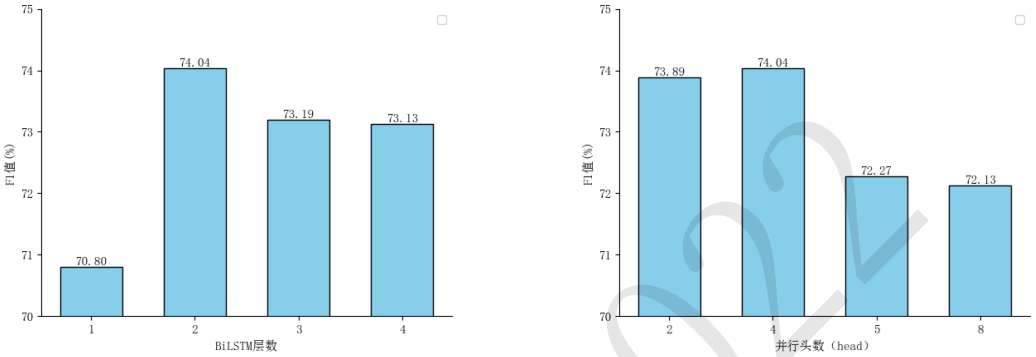
进一步地，从表 3中可知，本文所提模型效果均优于基线模型，尤其与现有因果关系抽取方法相比较，本文模型在准确率P、召回率R 和$F1$值上均有大幅提高，表明本文所提模型在中文因果关系抽取任务中，能够更加精确地抽取出文本中的因果事件对。另外，可以看到，本文

所提模型相较于基线模型在DUEE数据集上准确率的提升效果比CEC数据集显著。分析原因一是因为大规模数据能够减少噪声对因果关系抽取效果的影响，二是本文模型在大规模数据集上学习到的语义特征更为充分，能够更加准确地识别因果事件对。

此外，我们也发现加入预训练模型BERT后，使用BERT的模型效果比没有使用BERT的模型效果更好，表明预训练模型BERT丰富的语义知识能够提高因果关系抽取能力，并且BERT+本文模型的效果优于其他模型，表明本文提出的BiLSTM-TWAM+CRF模型能够有效地与预训练模型相结合，提高因果关系抽取性能。

4.5.2 超参数的选取

BiLSTM-TWAM层作为本文模型中的重要组成部分，为了评估BiLSTM的堆叠层数，以及TWAM中并行头(head)的个数对模型的影响，本文继续在CEC数据集上使用因果事件对作为评估标准进行了实验，实验结果如图 4所示。



(a) BiLSTM层数对$F1$值的影响　　(b) 并行头数(head)对$F1$值的影响

图 4: 模型参数设置对$F1$值的影响

从图 4可以看出，在其他参数相同的情况下，当BiLSTM层数与并行头数(head)分别取2和4时，模型抽取因果事件对的$F1$值最高。通过实验发现，将BiLSTM堆叠层数设置过大，会导致模型更加复杂和参数增多，学习到的特征会过于抽象而且会包含一些无用信息，设置过小又不能充分提取特征，使提取到的上下文特征缺乏语义信息表达。同时，当TWAM层中head数取5时，本文模型的$F1$值比head数取4时减少1.77%，表明随着head的数量增多，每个子空间的内含有的特征信息会逐渐减少，导致每个并行头无法提取到足够的特征信息，从而造成没有充足的语义特征表示。

4.5.3 消融实验

为了进一步验证本文所提BiLSTM-TWAM+CRF模型中每个组成部分的贡献，本文分别在CEC数据集和DUEE数据集上进行了消融实验，结果如表 4所示。

	CEC			DUEE		
模型	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$(%)
本文模型	76.42	**71.89**	**74.04**	**76.40**	**71.63**	**73.80**
-SoftLexicon	68.88	60.58	64.46	67.10	61.70	64.28
-TWAM	70.00	66.42	68.16	66.62	67.81	67.21
-残差融合	78.97	61.68	69.26	72.38	71.09	71.73

表 4: 消融实验结果

从表 4可以看出，本文所提模型的各个模块都发挥了一定的作用。模型在不使用SoftLexicon方法引入外部词汇信息时，CEC数据集与DUEE数据集的$F1$值分别下降了9.58%和9.52%，表明引入外部词汇信息能够增强字符向量的语义信息表示，使模型获取

到更多的文本特征信息。模型在不使用TWAM时，CEC数据集与DUEE数据集的$F1$值分别下降了5.88%和6.59%，表明引入TWAM能够在局部和全局两个角度学习文本序列中的特征信息，提取到更全面更深层次的语义特征。当没有采用残差结构融合TWAM和BiLSTM输出的特征时，CEC数据集与DUEE数据集的$F1$值分别下降了4.78% 和2.07%，表明该结构的应用，有效丰富了因果关系抽取任务中的语义特征，使得BiLSTM-TWAM层既能获取长距离的上下文特征表示，又能对特征在局部和全局角度进行更深层次的特征提取。实验结果表明，通过SoftLexicon方法引入的外部词汇信息对于模型的贡献最大，证明了在中文因果关系抽取任务中引入外部词汇信息的重要性，也表明基于字词联合的模型能够很大程度上提升中文因果关系抽取的能力。

5 结语

本文面向中文因果关系抽取，提出了一种基于外部词汇信息和注意力机制的因果关系抽取模型BiLSTM-TWAM+CRF，从一定程度上，实现了真正意义上的因果关系抽取。总体而言，本文提出的模型能够有效解决词语边界模糊和语义表征不充分的问题，具有较好的应用前景。后续工作将尝试从多特征融合角度来提升模型的多语种因果关系抽取效果。

参考文献

杨竣辉, 刘宗田, 刘炜, and 苏小英. 2016. 基于语义事件因果关系识别. 小型微型计算机系统, 37(3):433–437, January.

冯冲, 康丽琪, 石戈, and 黄河燕. 2018. 融合对抗学习的因果关系抽取. 自动化学报, 44(5):811–818, January.

苗佳, 段跃兴, 张月琴, and 张泽华. 2021a. 基于cnn-bigru模型的事件触发词抽取方法. 计算机工程, 47(9):69–74,83, September.

姜博, 左万利, and 王英. 2021b. 基于bert的因果关系抽取. 吉林大学学报*(*理学版*)*, 59(6):1439–1444, November.

郑巧夺, 吴贞东, and 邹俊颖. 2021c. 基于双层cnnbigrucrf的事件因果关系抽取. 计算机工程, 47(5):58–64,72, May.

Forney, G. D., and Jr. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Jianqi Gao, Xiangfeng Luo, Hao Wang, and Zijian Wang. 2021. Causal event extraction using iterated dilated convolutions with semantic convolutional filters. In *33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2021*, pages 619–623. IEEE, November.

Daniela Garcia. 1997. Coatis, an NLP system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management*, volume 1319 of *Lecture Notes in Computer Science*, pages 347–352. Springer, October.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, volume 12 of *MultiSumQA '03*, pages 76–83, USA. Association for Computational Linguistics.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005*, volume 4, pages 2047–2052.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014*, pages 987–997. Association for Computational Linguistics, June.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023.

第二十一届中国计算语言学大会论文集，第190页-第200页，南昌，中国，2022年10月14日至16日。

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *CoRR*, abs/1508.01991.

IAshwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Natural Language Processing and Information Systems*, volume 6716 of *Lecture Notes in Computer Science*, pages 52–63. Springer, June.

Devlin Jacob, Changming Wei, Lee Kenton, and Toutanova Kristina. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xianxian Jin, Xinzhi Wang, Xiangfeng Luo, Subin Huang, and Shengwei Gu. 2020. Inter-sentence and implicit causality extraction from chinese corpus. In *Advances in Knowledge Discovery and Data Mining*, volume 12084 of *Lecture Notes in Computer Science*, pages 739–751, Cham. Springer.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, May.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001),*, pages 282–289, San Francisco, CA, USA, June. Morgan Kaufmann.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2020. Simplify the usage of lexicon in chinese NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 5951–5960. Association for Computational Linguistics, July.

Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 909–918. Association for Computing Machinery, April.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg, and Dean Jeffrey. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pages 3111–3119, December.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, December.

Li Xinyu, Li Fayuan, Pan Lu, Chen Yuguang, Peng Weihua, Wang Quan, Lyu Yajuan, and Zhu Yong. 2020. Duee: A large-scale dataset for chinese event extraction in real-world scenarios. In *Natural Language Processing and Chinese Computing*, pages 534–545, Cham, October. Springer International Publishing.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344. ACL, August.

Ying Zeng, Honghui Yang, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2016. A convolution bilstm neural network model for chinese event extraction. In *Natural Language Understanding and Intelligent Applications*, volume 10102 of *Lecture Notes in Computer Science*, pages 275–287. Springer International Publishing.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pages 1554–1564, Melbourne, Australia, July. Association for Computational Linguistics.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China, October. ACL.

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950.