计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190042
57(1):159-174,2020

基于图注意力网络的因果关系抽取

许晶航[1] 左万利[1,2] 梁世宁[1] 王 英[1,2]
[1](吉林大学计算机科学与技术学院 长春 130012)
[2](符号计算与知识工程教育部重点实验室(吉林大学) 长春 130012)
(xujh17@mails.jlu.edu.cn)

Causal Relation Extraction Based on Graph Attention Networks

Xu Jinghang[1], Zuo Wanli[1,2], Liang Shining[1], and Wang Ying[1,2]
[1](*College of Computer Science and Technology, Jilin University, Changchun* 130012)
[2](*Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun* 130012)

Abstract Causality represents a kind of correlation between cause and effect, where the happening of cause will leads to the happening of effect. As the most important type of relationship between entities, causality plays a vital role in many fields such as automatic reasoning and scenario generation. Therefore, extracting causal relation becomes a basic task in natural language processing and text mining. Different from traditional text classification methods or relation extraction methods, this paper proposes a sequence labeling method to extract causal entity in text and identify direction of causality, without relying on feature engineering or causal background knowledge. The main contributions of this paper can be summarized as follows: 1) we extend syntactic dependency tree to the syntactic dependency graph, adopt graph attention networks in natural language processing, and introduce the concept of S-GAT (graph attention network based on syntactic dependency graph); 2) Bi-LSTM+CRF+S-GAT model for causal extraction is proposed, which generates causal label of each word in sentence based on input word vectors; 3) SemEval data set is modified and extended, and rules are defined to relabel experimental data with an aim of overcoming defects of the original labeling method. Extensive experiments are conducted on the expanded SemEval dataset, which shows that our model achieves 0.064 improvement over state-of-the-art model Bi-LSTM+CRF+self-ATT in terms of prediction accuracy.

Key words causal relation extraction; graph attention networks (GATs); sequence labeling; syntactic dependency graph; bidirectional long short-term memory (Bi-LSTM)

摘 要 因果关系作为一种重要的关系类型在关系推理等许多领域中起着至关重要的作用,因此对因果关系进行抽取是文本挖掘中的一项基本任务.与传统文本分类方法或关系抽取不同,采用序列标注的方法可以抽取文本中的因果实体并确定因果关系方向,不需要依赖特征工程或因果背景知识.主要贡献

收稿日期:2019-01-16;**修回日期**:2019-08-12
基金项目:国家自然科学基金项目(61976103,61872161);吉林省技术攻关项目(20190302029GX);吉林省自然科学基金项目(20180101330JC,20181013328JC);吉林省发改委项目(2019C053-8)
This work was supported by the National Natural Science Foundation of China (61976103, 61872161), the Project of Technical Tackle-key-problem of Jilin Province of China (20190302029GX), the Natural Science Foundation of Jilin Province of China (20180101330JC, 20181013328JC), and the Project of the Development and Reform Commission of Jilin Province (2019C053-8).
通信作者:左万利(zuowl@jlu.edu.cn)

有:1)拓展句法依存树到句法依存图,将图注意力网络应用到自然语言处理中,引入了基于句法依存图的图注意力网络的概念;2)提出 Bi-LSTM+CRF+S-GAT 因果关系抽取模型,根据输入的词向量生成句子中每个词的因果标签;3)对 SemEval 数据集进行修正与拓展,针对其存在的缺陷制定规则重新标注实验数据.在拓展后的 SemEval 数据集上进行了大量的实验,结果表明:该模型在预测准确率上比现有最优模型 Bi-LSTM+CRF+self-ATT 提高了 0.064.

关键词 因果关系抽取;图注意力网络;序列标注;句法依存图;双向长短期记忆网络

中图法分类号 TP391

因果关系,即"原因"与"结果"之间的对应关系,是一种重要的关系类型,在事件检测和预测[1]、回答问题[2]、情景生成[3]等任务中起着十分重要的作用.因此,因果关系抽取是文本挖掘中的一项基本任务.自然语言文本或日常的信息交流中均存在大量的因果关系,例如新闻中报道交通状况:"〈e1〉Accident〈/e1〉 causes 〈e2〉delays〈/e2〉 on bw parkway.",标签〈e1〉内的单词"accident"表示原因,标签〈e2〉内的单词"delays"表示结果,即"事故"导致了"堵塞".

根据原因与结果之间的对应关系可以将因果关系分为一因一果、一因多果、多因一果与多因多果.一因一果指 1 个原因与 1 个结果相对应,一因多果意为存在 1 个原因致使多个结果的发生,多因一果的含义为多个原因共同导致 1 个结果,多因多果则表示文本中存在多个原因共同导致了多个结果.还可以根据是否含有因果连接词分为显式因果关系与隐式因果关系.显式因果关系指原因与结果之间有特定的语言成分连接,例如中文中的"导致"、"引起"等,英文中的"cause","result in"等.隐式因果关系常表现为单句内不含因果连接词,或因果分布在不同句中.

当前针对因果关系抽取的研究主要包括 3 种类型.

1) 文本分类.判断句子是否含有因果关系.

2) 关系抽取.给定句中的因果候选对,判断是否具有因果关系.

3) 序列标注.对句子进行序列标注,抽取因果实体并确定因果关系方向.

主流方法虽对因果关系进行了多角度研究,但仍存在 3 点局限:

1) 通常基于大量特征以提高模型能力,但构建特征的自然语言处理(natural language processing, NLP)工具由于自身缺陷会导致错误传播,且特征选择过程十分繁琐复杂;需要大量的 NLP 工作,如词性标注、语义分析等;需要因果关系背景知识,如从大量的语料库中总结出潜在含有因果关系的实体(例如"地震"与"死亡",发生地震通常会导致人员的死亡)来提高因果关系抽取的准确率.

2) 现有方法多是根据句子是否含有因果关系进行分类,将问题转化为文本分类问题;或根据给定的候选对判断是否具有因果关系,问题则被简化为关系抽取问题.二者均未能做到真正的"抽取".

3) 多数工作将因果数量限定为一因一果,未对多因多果的句式进行探究;依赖因果连接词,只能抽取带标记的显式因果关系;限于句内因果,无法探究跨句、跨段的因果关系.

本文采取序列标注的方法抽取因果关系.随着注意力机制的火热,许多序列标注模型通过引入注意力的思想提高模型的准确度.传统序列标注模型中的注意力机制是在处理局部信息时同时关注整体的信息,数据是线性且互相独立的,但实际的句子中词之间存在很强的依赖关系.对句子进行句法依存分析得到句法依存树,通过树中节点间的依存弧获取词间的依赖关系.但树形结构存在一定的局限性(例如弧是单向的),故本文将句法依存树拓展到句法依存图,使传统注意力机制中互相独立的线性数据转化为具有依赖关系的图形数据.应用图注意力网络(graph attention network, GAT)[4]到 NLP 中,引入了基于句法依存图的图注意力网络(graph attention network based on syntactic dependency graph, S-GAT),并提出了 Bi-LSTM+CRF+S-GAT 因果关系抽取模型.模型在识别句中词的因果标签时,对词在句法依存图中的相邻节点进行注意力计算,为其分配不同的权重,从而关注作用较大的词,忽视作用较小的词,增加因果抽取的准确度.

本文主要贡献有 3 个方面:

1) 拓展句法依存树到句法依存图,将 GAT 应用到 NLP 中,提出了 S-GAT.句法依存图中的顶点对应句中的词,图中的边表示词之间的依赖关系,使传统注意力机制中的线性数据转化为图形数据.在

计算注意力时,根据句法依存图中相邻节点更新权重,使注意力更集中在表示“原因”和“结果”含义的词上.

2) 提出 Bi-LSTM+CRF+S-GAT 因果关系抽取模型,根据输入的词向量生成句子中每个词的因果标签.将传统用在词性标注、命名实体识别等领域的序列标注方法应用在因果关系抽取中,标注出因果实体并确定因果关系方向,实现了真正的“抽取”.将一因一果拓展到多因多果,不限于显式因果关系,不需要构建特征工程,不需要繁琐的 NLP 工作和因果关系背景知识.

3)对 SemEval 数据集进行修正与拓展,针对其存在的缺陷制定规则重新标注实验数据.在拓展后的 SemEval 数据集上进行实验,结果表明,本文提出的因果关系抽取模型 Bi-LSTM+CRF+S-GAT 较现有模型的因果抽取效果有明显的提升,相比 Bi-LSTM+CRF+self-ATT 模型的准确率提高了 0.064.

1 相关工作

1.1 因果关系抽取

因果关系的抽取方法主要分为 2 种:基于模式匹配方法和基于机器学习方法.

1.1.1 基于模式匹配方法

文献[5]分析了法语中具有因果含义的动词,实现了 1 个 COATIS 系统用于抽取含有“Cause Verb Effect”结构的带有标记的显示因果关系.文献[6]认为除动词之外,还有些介词(“for”,“from”等)、状语连接词(“so”,“hence”,“therefore”等)以及子句(“that's why”,“the result is”等)也能表达因果关系,采用模式匹配方法从人工标注的 Wall Street Journal 语料中抽取带标记的因果关系.

1.1.2 基于机器学习方法

现阶段基于机器学习或深度学习模型对因果关系抽取的研究主要包括 3 类.

1) 文本分类.对句子是否含有因果关系进行分类.文献[7]提出了 2 种方法:①基于知识特征的分类模型;②基于深度学习的方法,用卷积神经网络(convolutional neural network, CNN)对因果关系进行分类.该模型能够识别显式因果和隐式因果以及因果关系的方向.文献[8]用平行的维基百科语料库识别新的标记(已知的因果短语的变体),通过远程监督创建训练集,使用开放类标记的特征与上下文信息的语义特征训练因果分类器.

2) 关系抽取.判断给定的候选对是否含有因果关系.文献[9]将 SemEval 数据集中的单词扩展到短语,一因一果拓展到多因多果,提出了一种新的约束隐藏朴素贝叶斯模型提取文本中的显式因果关系.文献[10]用生成式对抗网络(generative adversarial networks, GANs)的对抗学习特性,将带有注意力机制的双向门控循环单元网络(bidirectional gated recurrent units networks, BGRU)与对抗学习相融合,提出了一种融合对抗学习的因果关系抽取方法,避免了繁琐的特征工程.文献[11]用多列卷积神经网络抽取因果关系,使用了从网络文本中提取的背景知识以及从原始句子中提取的因果关系候选信息,需要大量的 NLP 预处理工作.

3) 序列标注.标注出因果实体.文献[12]用层叠条件随机场对事件间的因果关系进行抽取,将因果关系扩展到跨句、跨段、多因多果等多种类型,人工构建大量的特征工程.文献[13]利用单词级别的词向量及其语义特征,通过双向长短期记忆网络(bidirectional long short-term memory, Bi-LSTM)标注出句子中的原因、结果以及因果连接词,并将标记的单词扩展到短语(包括虚词“of”等).文献[14]利用因果关系的时间特性,将因果抽取重新定义为一种特殊的时间提取方法,通过引入多层条件随机场模型将任务转化成一个序列标注的过程.

此外,文献[15]针对告警关联分析中因果知识难以自动获取的问题,利用 Markov 链模型(Markov chain model, MCM)对因果知识进行建模,提出了一种基于 Markov 性质的因果知识挖掘方法.文献[16]将充分必要条件与因果关系融合,引入了充分因果关系与必要因果关系的概念,提出一种模拟术语之间因果关系强度的度量方法,探究了短文本(短语和句子)之间常识性因果关系的推理问题.

早期研究根据因果句子的结构特点进行模式匹配抽取因果关系,该方法只能抽取固定模式且带标记的显式因果关系.随着机器学习理论的逐渐成熟,因果抽取的范围逐渐扩大,拓展到多因多果、隐式因果、跨句、跨段等多种形式的因果关系.支持向量机(support vector machine, SVM)、朴素贝叶斯模型(naive Bayesian model, NBM)等机器学习算法可以根据句子或句中的候选对是否含有因果关系进行分类,条件随机场等序列标注模型能够标注出因果实体并确定因果关系方向.近些年随着深度学习的火热,卷积神经网络、循环神经网络(recurrent neural network, RNN)等深层网络结构被加入到因果关系

抽取的模型中,与传统的机器学习算法相比,基于深度学习的因果关系抽取模型的效果有明显提升.由于文本分类与关系抽取的方法均未做到真正的因果"抽取",本文采用序列标注结合深度学习中的 Bi-LSTM 网络与注意力机制抽取因果关系.

1.2 序列标注与注意力机制

序列标注是解决 NLP 问题时经常遇到的基本问题之一,传统方法有隐 Markov 模型(hidden Markov model, HMM)、条件随机场(conditional random field, CRF)等.Lafferty 等人[17]提出的 CRF 是一种无向图模型,结合了最大熵模型与隐 Markov 模型的特点,在分词、词性标注等序列标注任务中取得了很好的效果.

长短期记忆网络(long short-term memory, LSTM)有时序性和长依赖的特点,Bi-LSTM 能够获取跨度较远的过去与将来时序的数据特征,挖掘更丰富的语义信息.结合 CRF,文献[18]提出了 Bi-LSTM+CRF 模型,使序列标注的效果有了显著的提升.

Google 从自然语言本身的特性出发,提出了一种完全基于注意力机制[19]的网络框架 Transformer.结合深度学习模型与注意力机制,文献[20]提出了 CNN/RNN/FNN(feed-forward neural network)+self-ATT 模型进行语义角色标注.

图卷积(graph convolutional networks, GCNs)[21]是一种用可扩展的方法对图形结构数据进行半监督学习的网络结构,是基于直接对图进行操作的卷积神经网络的一种有效变体,通过谱图卷积的一阶近似局部化来促进卷积体系结构的选择.结合图卷积与注意力机制,文献[4]提出了图注意力网络.GAT 是基于图形结构化数据的新型神经网络,根据相邻节点进行权重更新,利用掩码的自注意力层解决图卷积或与其近似的现有方法的缺点.GAT 的提出,将注意力机制中的线性数据转化为图形数据,使原本互相独立的数据通过"图"产生联系.

本文采用序列标注抽取因果关系,在传统序列标注模型与注意力机制的基础上,结合 GAT 提出了 Bi-LSTM+CRF+S-GAT 模型.

2 传统序列标注 Bi-LSTM+CRF 模型

2.1 条件随机场(CRF)

CRF 是给定随机变量 X 条件下,计算随机变量 Y 的 Markov 随机场.

$$P(Y\mid X)=\frac{1}{Z_X}\exp\Big(\sum_{s=1}^{S}\sum_{l}\lambda_l f_l(y_{s-1},y_s,X,s)\Big),$$

$$Z_X=\sum_{y\in Y}\exp\Big(\sum_{s=1}^{S}\sum_{l}\lambda_l f_l(y_{s-1},y_s,X,s)\Big), \tag{1}$$

$P(Y|X)$表示当前时刻标记转移 $y_{s-1}\to y_s$ 与输入序列 X 的任意特征值.其中 X 是输入变量,表示观测序列;Y 是输出变量,表示标记序列;Z_X 为归一化因子;$f_l(y_{s-1},y_s,X,s)$是特征函数.训练时通过极大似然估计得到条件概率模型,并用该模型进行预测.

CRF 在词性标注、命名实体识别等序列标注任务中取得了很好的效果.以词性标注为例,观测序列为词序列,输出的标签为每个词对应的词性.当前词与相邻上一词间满足一定条件时,其特征函数 $f_l=1$,否则 $f_l=0$.每个词都用相同个数的特征函数进行判断,是全局最优化值.预测的过程是用每种特征配置给标签打分并加权求和,得分最高的标签就是预测的结果.

2.2 双向长短期记忆网络(Bi-LSTM)

CRF 可以通过获取相邻词的特征预测当前词的标签,但在一些序列标注的任务中,需要通过获取长距离特征来提高序列标注的准确度.

2.2.1 循环神经网络(RNN)

RNN 允许信息的持久化,利用其内部记忆处理带有时序性的数据.含有因果关系的句子可以理解为 1 个时序性的序列,每个词对应 1 个时间的数据.在处理当前词时,RNN 能够获取过去时序词的特征.

2.2.2 长短期记忆网络(LSTM)

对于时序间隔较长的数据,RNN 难以获取句子中跨度较远词的特征,而 LSTM 解决了这种长期依赖的问题.图 1 为 LSTM 的细胞结构图.

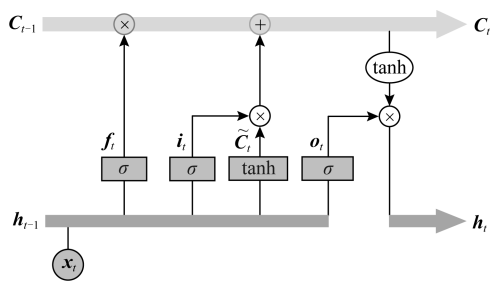


Fig. 1 LSTM cellular structure

图 1 LSTM 细胞结构

首先,当前输入信息 $\boldsymbol{x}_t$ 与上一时刻隐藏层输出 $\boldsymbol{h}_{t-1}$ 通过函数 *sigmod* 得到从遗忘门中丢弃的信息 $\boldsymbol{f}_t$.

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f). \tag{2}$$

然后,根据需要更新的信息 $\boldsymbol{i}_t$ 与生成备选更新的内容 $\widetilde{\boldsymbol{C}}_t$ 通过输入门将旧细胞状态 $\boldsymbol{C}_{t-1}$ 更新为新细胞状态 $\boldsymbol{C}_t$.

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i), \tag{3}$$

$$\widetilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C), \tag{4}$$

$$\boldsymbol{C}_t=\boldsymbol{f}_t\cdot\boldsymbol{C}_{t-1}+\boldsymbol{i}_t\cdot\widetilde{\boldsymbol{C}}_t. \tag{5}$$

最后,输入信息 $\boldsymbol{x}_t$ 与上一时刻隐藏层输出 $\boldsymbol{h}_{t-1}$ 通过函数 *sigmod* 得到输出信息 $\boldsymbol{o}_t$,新细胞状态 $\boldsymbol{C}_t$ 通过 tanh 层与 $\boldsymbol{o}_t$ 相乘得到当前时刻隐藏层的输出 $\boldsymbol{h}_t$.

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o), \tag{6}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\cdot\tanh(\boldsymbol{C}_t). \tag{7}$$

2.2.3 双向长短期记忆网络(Bi-LSTM)

LSTM 能够获取过去时序数据的特征,但在一些序列标注任务中还需要考虑将来时序的数据.Bi-LSTM 由前向的 LSTM 与后向的 LSTM 组成,可以同时获取过去和将来时序的数据特征,从而获取上下文信息,其结构如图 2 所示:

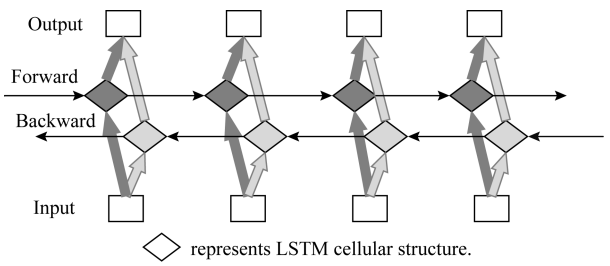


Fig. 2 Bi-LSTM structure

图 2 Bi-LSTM 结构

2.3 Bi-LSTM+CRF 模型

结合 Bi-LSTM 与 CRF 模型的结构特点,文献[18]提出了 Bi-LSTM+CRF 模型进行序列标注,其结构如图 3 所示.首先对输入的句子进行向量化表示,再通过 Bi-LSTM 层获取过去与将来的输入信息并自动提取句子特征,最后接入 CRF 层进行句子级别的标签预测.该模型能够有效地根据过去与将来的标签信息预测当前数据的标签.

以实体命名识别为例,如图 3 所示,输入为句子“EU rejects German call”,输出为单词对应的实体标签.其中标签“ORG”表示“组织名”,“MISC”表示“杂项”,“O”表示“非实体”;“B”是用于划分实体的边界,表示实体标签的开始.该模型识别出“EU”为组织名,“German”为杂项,意为其他实体.

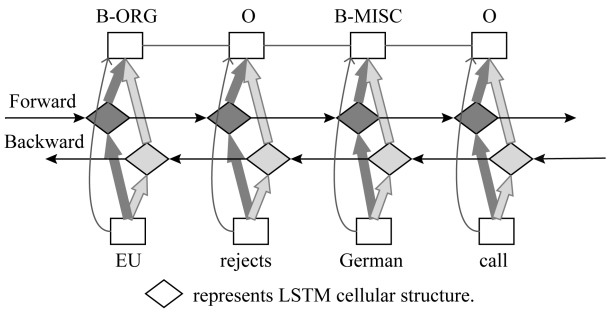


Fig. 3 Bi-LSTM+CRF model structure

图 3 Bi-LSTM+CRF 模型结构

该模型在序列标注相关应用中的效果优于 CRF,LSTM,Bi-LSTM 等模型,因此 Bi-LSTM+CRF 成为了序列标注的基本模型,现有的序列标注工作大多是在该模型的基础上进行改进.

3 Bi-LSTM+CRF+S-GAT 模型

3.1 图注意力网络(GAT)

GAT[4]是将图卷积和注意力机制结合的网络结构,在输入待处理的图自身上计算注意力.每一个节点更新隐藏层输出时,都要对其相邻节点进行注意力计算,目的是为每个相邻节点分配不同的权重,从而关注作用较大的节点,忽略作用较小的节点.其特点是计算高效,每一个节点与其相邻节点计算注意力是并行的,且可以分配任意的权重给相邻节点.

3.2 句法依存图

句法依存分析是根据句子中词与词之间的依存关系表示词语的句法结构信息(如主谓、动宾、定中等结构关系),并用树状结构表示句子结构(如主谓宾、定状补等)的一种 NLP 关键技术.以句子“flu causes cold.”为例,对句子进行句法依存分析,结果如图 4 所示.根节点(Root)为单词“causes”,“causes”分别为“flu”,“cold”,“.”的父节点,“Nsubj”表示“flu”是“causes”的名词性主语.

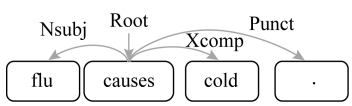


Fig. 4 Syntactic dependency analysis

图 4 句法依存分析

传统的句法依存分析以树形结构存储,只能从父节点指向子节点,即依赖关系是单向的,且节点本身

不能指向自己.针对树形结构存在的缺陷,本文将句法依存树拓展到图并提出句法依存图的概念.句法依存图的生成规则为:句中的词为句法依存图的顶点,根据句法分析得到的句法依存树的弧生成句法依存图的边.忽略"Root"指向根节点的弧,其他句法依存树的弧为句法依存图的边.由于句法依存图注重词之间的依赖关系,不关注句法结构(如主谓等结构关系),故不需要存储句法依存树中弧的标签信息(如"Nsubj"等).

本文以邻接矩阵的方式存储句法依存图,有边的对应矩阵元素为1,否则为0.句法依存图分为3类:

1) 有向图.句法依存分析中的父节点指向子节点,弧是单向的.

2) 无向图.句法依存分析中父节点指向子节点,同时子节点也指向父节点,父子节点间有一个无向弧连接,邻接矩阵为对称矩阵.

3) 带有"self-arc"的图.节点自身指向自己,邻接矩阵对角元素为1.

常用的句法依存图有:有向图、无向图、有向图+"self-arc"、无向图+"self-arc".根据图4所示的句法依存树生成相应的句法依存图,其邻接矩阵如图5所示.以生成无向图+"self-arc"的句法依存图为例,图的顶点分别为句中单词"flu","causes","cold",".";忽略句法依存树中的弧"Root→causes",将其他的弧"causes→flu","causes→cold","causes→."转换成无向图的边,分别为正向边"causes-flu","causes-cold","causes-.",以及反向边"flu-causes","cold-causes",".-causes";此外,带有"self-arc"图的边分别为"flu-flu","causes-causes","cold-cold",".-.".所有的顶点与边构成无向图+"self-arc"的句法依存图,其邻接矩阵如图5(d)所示,是一个对角元素为1的对称矩阵.

	flu	causes	cold	.
flu	0	0	0	0
causes	1	0	1	1
cold	0	0	0	0
.	0	0	0	0

(a) Directed graph

	flu	causes	cold	.
flu	0	1	0	0
causes	1	0	1	1
cold	0	1	0	0
.	0	1	0	0

(b) Undirected graph

	flu	causes	cold	.
flu	1	0	0	0
causes	1	1	1	1
cold	0	0	1	0
.	0	0	0	1

(c) Directed graph+"self-arc"

	flu	causes	cold	.
flu	1	1	0	0
causes	1	1	1	1
cold	0	1	1	0
.	0	1	0	1

(d) Undirected graph+"self-arc"

Fig. 5 The adjacency matrix of the syntactic dependency graph

图5 句法依存图的邻接矩阵

3.3 基于句法依存图的图注意力网络(S-GAT)

结合句法依存图与图注意力网络,本文提出了S-GAT.S-GAT中图的顶点与句中的词相对应,顶点的特征为词的向量特征,图的边对应句法依存图的边.

3.3.1 强化特征

$\boldsymbol{h}=(\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_I)$,$\boldsymbol{h}_i\in\mathbb{R}^F$,其中$\boldsymbol{h}$为句子对应的向量,在本文模型中为词向量输入到Bi-LSTM的隐藏层输出.I为句子分词后词的个数,F为隐藏层输出的特征维度(词向量维度).为了得到表达能力更强的Bi-LSTM隐藏层输出,用1个可学习的线性变换将隐藏层输出特征转化为更高层次的特征,将$\boldsymbol{W}\in\mathbb{R}^{F\times F'}$的权重矩阵作用到Bi-LSTM的隐藏层输出上得到强化后的特征.句中词的个数I保持不变,改变了隐藏层输出的特征维度F'值.

3.3.2 计算注意力系数

得到强化的特征后,对每个词进行自注意力(共享注意力机制$a:\mathbb{R}^F\times\mathbb{R}^{F'}\to\mathbb{R}$)计算.

$$e_{ij}=a(\boldsymbol{W}\boldsymbol{h}_i,\boldsymbol{W}\boldsymbol{h}_j)=LeakyReLU(\boldsymbol{W}_a^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\,\|\,\boldsymbol{W}\boldsymbol{h}_j]),\tag{8}$$

e_{ij}表示词j对于词i的重要程度.注意力机制a是1个单层前馈网络,将权重矩阵$\boldsymbol{W}_a\in\mathbb{R}^{2F'}$作用在强化后的特征上;"$\|$"表示连接.如图6所示,将词$i$、词$j$强化后的特征向量相连接,输入到单层前馈网络,通过函数$LeakyReLU$非线性层得到e_{ij}.

$$\alpha_{ij}=softmax_j(e_{ij})=\frac{\exp(e_{ij})}{\sum\limits_{k\in N_i}\exp(e_{ik})}=\frac{\exp(LeakyReLU(\boldsymbol{W}_a^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\,\|\,\boldsymbol{W}\boldsymbol{h}_j]))}{\sum\limits_{k\in N_i}\exp(LeakyReLU(\boldsymbol{W}_a^{\mathrm{T}}[\boldsymbol{W}\boldsymbol{h}_i\,\|\,\boldsymbol{W}\boldsymbol{h}_k]))},\tag{9}$$

式(9)计算卷积时用来加权求和的系数α_{ij}.设词i在句法依存图中直接相连的节点集合为N_i,e_{ij}通过函数$softmax_j$得到注意力系数α_{ij}.

将词i在句法依存图中所有相邻词j的强化特征与对应的权重系数α_{ij}进行加权求和,通过非线性层σ得到注意力特征$\boldsymbol{h}_i'$:

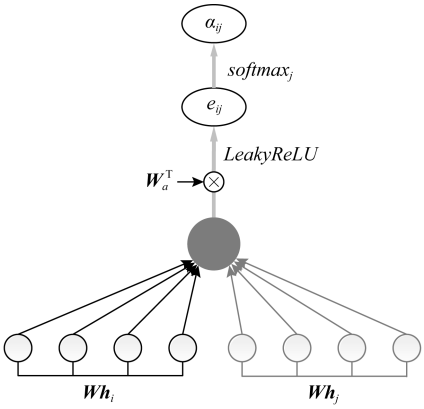


Fig. 6 Attention coefficients

图 6 注意力系数

$$\boldsymbol{h}'_i = \sigma\Big(\sum_{j\in N_i}\alpha_{ij}\boldsymbol{W}\boldsymbol{h}_j\Big). \tag{10}$$

3.3.3 多头注意力机制

为了使模型结构更稳定,将 3.3.2 节方法拓展到多头注意力机制(multi-head attention mechanism)[19].

$$\boldsymbol{h}'_i = \Big\|_{k=1}^{K}\sigma\Big(\sum_{j\in N_i}\alpha_{ij}^k\boldsymbol{W}^k\boldsymbol{h}_j\Big), \tag{11}$$

其中,K 为头的个数,$\boldsymbol{W}^k$ 为强化特征的权重矩阵,α_{ij}^k 是计算 k-th 注意力的权重系数.K 个互相独立的注意力机制按式(11)所示的方法进行变换,连接其特征得到输出 $\boldsymbol{h}'_i$.S-GAT 的最后 1 层(输出层),选择

$$\boldsymbol{h}'_i = \sigma\Big(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in N_i}\alpha_{ij}^k\boldsymbol{W}^k\boldsymbol{h}_j\Big), \tag{12}$$

对注意力特征取平均值得到输出结果.

以 $K=3$ 为例,如图 7 所示,$\boldsymbol{h}_2\sim\boldsymbol{h}_5$ 为 $\boldsymbol{h}_1$ 的相邻节点的向量(包含 $\boldsymbol{h}_1$ 自身),3 种不同的箭头代表 3 种互相独立的注意力计算,用连接或取平均值的方法得到输出结果 $\boldsymbol{h}'_i$.

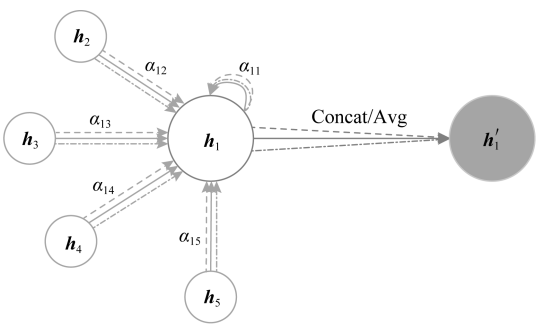


Fig. 7 Multi-head attention

图 7 多头注意力

3.3.4 复杂度分析

S-GAT 的时间复杂度为 $O(|\boldsymbol{V}|FF'+|\boldsymbol{E}|F')$,其中 F 为 Bi-LSTM 隐藏层输出的特征维度(词向量维度),F'为加强后的特征维度,$|\boldsymbol{V}|$是句法依存图中顶点的个数(即分词后词的个数 I),$|\boldsymbol{E}|$是句法依存图中边的个数.S-GAT 时间复杂度的计算分为 2 部分:计算注意力系数的时间复杂度与计算卷积的时间复杂度.

计算注意力系数时,每个词计算注意力的时间复杂度为 $O(FF')$,且只计算与其在句法依存图中直接相连的词的系数,即 $O(|\boldsymbol{V}|)$,因此计算注意力系数的时间复杂度为 $O(|\boldsymbol{V}|FF')$.

计算卷积时,每个句法依存图的边均对应 1 个词的 Bi-LSTM 隐藏层输出乘上权重并包含到卷积的加权求和中,强化后的特征维度为 F',因此计算卷积的时间复杂度为 $O(|\boldsymbol{E}|F')$.

3.4 Bi-LSTM+CRF+S-GAT 模型

本文结合序列标注传统模型 Bi-LSTM+CRF 与 S-GAT 提出了 Bi-LSTM+CRF+S-GAT 模型,模型结构如图 8 所示,将 LSTM Layer+Linear Layer+S-GAT Layer 设为 1 个块(block),堆叠 N 层.模型根据输入的句子输出句中词对应的因果标签.

1) 输入层.输入含有因果关系的句子,将句子分别输入到神经网络层提取特征,以及句法分析层生成句法依存图.

2) 神经网络层.利用训练好的 GloVe 模型对分词后的句子进行向量化表示,将输入的语言文字转化为特征向量,并通过 Bi-LSTM 网络挖掘语义信息,充分利用上下文提取句子的深层语义特征,进而探究语义中潜在的因果关系.

3) 句法分析层.对输入的句子进行句法分析,得到句法依存树,根据 3.2 节中生成句法依存图的方法生成相应的邻接矩阵.图 8 中的顶点与句子中的单词相对应,单词间的依存弧为图的边,线性数据转换为图形数据.

4) S-GAT 层.将 Bi-LSTM 隐藏层输出通过线性层转化为更高层次的特征,获取表达能力更强的隐藏层输出.将强化后的特征与句法分析层生成的邻接矩阵输入到融合层,进行图注意力的计算.利用式(8)(9)计算注意力系数,并根据式(10)进行加权求和得到注意力特征.为了使模型结构更稳定,S-GAT 采用多头注意力机制的思想,将 K 个互相独立的注意力特征根据式(11)进行连接得到 1 层 S-GAT 的输出,该层堆叠 n 次.S-GAT 的输出层是根据式(12)将注意力特征取平均值得到 S-GAT 的

最终结果,S-GAT 层数设为 n'.S-GAT 将强化后的线性特征转化为图形特征,原本互相独立的词特征通过句法依存图的边产生依赖关系.每个单词在计算自身注意力时,为其所有相邻单词分配不同的权重,进而关注作用较大的单词,忽略作用较小的单词,使注意力更集中在要抽取的原因词"flu"和结果词"cold"上,进一步加强了因果语义的特征.

5) CRF 层.将 S-GAT 层的输出通过 CRF 层得到最终的因果标签.CRF 能够获取相邻词的信息,用多种特征函数给标签打分并加权求和,得分最高的标签为最终的输出结果,是全局最优化值.

6) 输出层.最终的标签结果如图 8 所示,其中标签"C"代表"cause"表示原因,"E"代表"effect"表示结果,"O"代表"other"表示无因果关系.

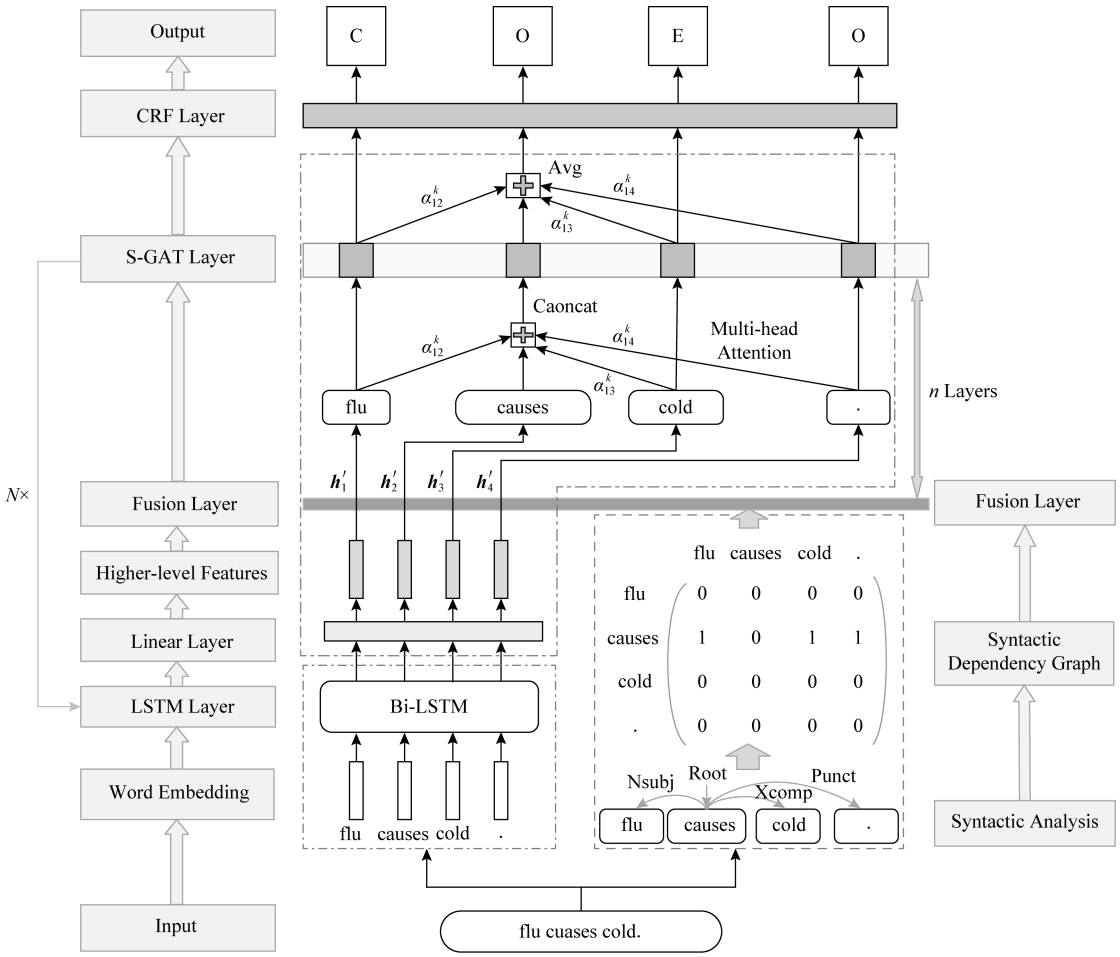


Fig. 8 Bi-LSTM+CRF+S-GAT model

图 8 Bi-LS TM+CRF+S-GAT 模型

模型输入为句子"flu causes cold.".首先将输入的语言文字转化为特征向量,并通过 Bi-LSTM 挖掘上下文的语义信息,初步提取句子的因果语义特征;同时对输入的句子进行句法分析得到句法依存图,使句子中原本互相独立的单词之间产生依赖关系.然后,将初步提取的词特征根据依赖关系通过 S-GAT 分配权重,进一步强化因果语义特征.最后,通过 CRF 层提取近距离的数据特征,并输出因果标签.得到的因果关系的抽取结果为:原因为"flu",结果为"cold",即"flu"导致了"cold",达到了因果抽取的目的.

4 数据与实验

4.1 实验数据

4.1.1 数据来源

实验数据来自 SemEval 数据集与英文维基百科语料库.本文选取关系分类数据集 SemEval (SemEval 2007 Task4 与 SemEval 2010 Task8)中含有因果关系的 1 368 个句子作为实验数据.该数据量较少,不足以满足实验需求,故需要对实验数据进行扩充.文献[8]从英文维基百科语料库中提取了含有因果关

系的句子,本文从该数据集中挑选出 1 632 个句子进行人工标注,标注的标准与 SemEval 数据集一致,最终得到了含有因果关系的 3 000 个句子作为实验数据.

SemEval 数据集原始数据的一个例子:

① 011 "〈e1〉Zinc〈/e1〉 is essential for 〈e2〉growth 〈/e2〉 and cell division."

② *WordNet*(e1) = "Zinc% 1:27:00::", *WordNet*(e2) = "growth% 1:22:00::", *Cause-Effect*(e1,e2) = "true", Query = "* is for growth"

第 1 行"011"代表句子序号,引号内是含有因果关系的句子,标签〈e1〉〈/e1〉,〈e2〉〈/e2〉内的单词表示原因或结果;第 2 行中 *Cause-Effect*(e1,e2) = "true"表示句子含有因果关系,且因果关系的方向是 e1→e2,即"Zinc"是原因,"growth"是结果.

4.1.2 数据修正

SemEval 是关系分类的数据集,用来识别给出候选对所具有的关系.该数据集只关注标签内的词是否具有因果关系,忽略了标签外的词.原始数据集 SemEval 存在的问题有:

1) 存在未标记出的因果关系

原数据集中的句子只有一组候选对,故对其进行标注只限于一因一果,但很多句子存在多因多果.例如句子"〈e1〉Frustrations〈/e1〉, threats, and conflicts cause 〈e2〉stress〈/e2〉.","frustrations"为原因,"stress"为结果.由语义可知,"threats","conflicts"也是"stress"的原因,但原数据集没有进行标记.

2) 标注长度不一致

例句 1:"Mr c notes 〈e1〉worsening〈/e1〉 of seizures and cognition with sleep 〈e2〉deprivation 〈/e2〉 and stress.";

例句 2:"Mr c notes 〈e1〉 worsening〈/e1〉 of seizures and cognition with 〈e2〉sleep deprivation 〈/e2〉 and stress.".

例句 1 中〈e2〉标签内为单词"deprivation",例句 2 中为短语"sleep deprivation",标注单词还是短语存在歧义.

3) 连锁因果关系

例如句子"The aircraft was written off in the accident due to the severe impact caused by the KLM aircraft,and the resulting fire.",由语义可知,"impact"是"accident"的原因,同时也是"aircraft","fire"的结果,"impact"如何标注存在歧义.

4) 包含性歧义

在英文中,存在例如"including","such as"等具有"包含"语义的词会影响因果关系的标注.例如句子"Most illnesses, including colds and 〈e1〉flu 〈/e1〉,cause a toxic overload that also increases the 〈e2〉stress〈/e2〉 on the kidneys.",句中的"illness"是"colds"和"flu"的统称,"colds"和"flu"是"illness"的特例,标记统称"illness"还是特例"colds","flu"存在争议.且"colds"和"flu"是并列关系,原数据集只标记了"flu".

5) "of"短语歧义

在英文中一些带有"of"的短语在语义上是一个整体,由于 SemEval 数据集只标记实体单词,故只标记了"of"短语的一部分.例如句子"A first 〈e1〉revolution〈/e1〉 was triggered by the growing use of 〈e2〉reading〈/e2〉 and writing",在短语"use of reading and writing"中标记"use"还是"reading","writing"存在歧义.

数据集中还存在其他问题,本文不全部列出.

针对上述 5 个问题本文对原数据集进行了修正,规则为:

1) 人工标注出原数据集未标记的因果关系实体(例如多因多果).

2) 短语统一选取核心单词进行标注(1 个单词).

3) 连锁因果中既是原因又是结果的单词与其后面因果单词的标记相同.

4) 包含性歧义统一标记"统称"的单词,忽略"特例"的单词.

5) 带有"of"的短语,在不影响语义的情况下优先标记"of"前的单词.

数据集中其他问题也做了相应的修正,在标记存在争议时采用投票表决的方法决定最终的标签.英文维基百科数据集的标注标准与上述规则完全一致.

4.1.3 标注方法

本文采用序列标注进行因果关系抽取,故需要给句子中的每个词标记相应的标签.为了简单明了地将因果关系用标注的方法抽取出来,本文选取"C","E","O"这 3 种标签进行标注,其中"C"代表"cause"表示原因,"E"代表"effect"表示结果,"O"代表"other"表示该词不具有因果关系.

如 4.1.2 节所述,本文的原因与结果均由核心单词(1 个单词)表示,故对于一因一果,标注序列中含有 1 个标签"C"和 1 个标签"E",一因多果则含有

1个标签“C”和多个标签“E”,多因一果的标注序列存在多个标签“C”和1个标签“E”,有多个标签“C”和“E”的标注序列表示该句子含有多因多果.

由于本文涉及句法依存分析,故标点符号也同单词一样进行标注(标注为“O”).由于不标注因果连接词,本文的因果关系抽取不限于带有标记的显式因果,标注例子如表1所示:

Table 1 Example of Labeling

表1 标注例子

Word	Label
Frustrations	C
,	O
threats	C
and	O
conflicts	C
cause	O
stress	E
.	O

4.1.4 显式因果与隐式因果

在因果关系抽取中,根据是否含有因果连接词将句子分为显式因果与隐式因果.根据文献[22]的显隐式因果分类方法对数据进行分类.

1) 显式因果

① 显式连接词,如“cause”,“result in”等具有明显因果含义的动词.例句:The 〈e1〉pollution 〈/e1〉 was caused by the 〈e2〉shipwreck〈/e2〉.

② 模糊连接词:无明显的因果含义.

(i) 可以通过结果性和工具性的动词模式来实现因果含义的连接词,如“increase”,“trigger”等.例句:The Maze procedure crates new pathways for the electrical 〈e1〉impulses〈/e1〉 that trigger the 〈e2〉 heartbeat〈/e2〉.

(ii) 使因果代理与产生情况不可分割的连接词,如“plague(by)”,“generate(by)”等.例句:The noise 〈e1〉signal〈/e1〉 was generated by a noise 〈e2〉 diode 〈/e2〉 (ENR=27 dB) and gated with a high-speed electronic switch.

(iii) 非动词模式,如“due to”,“from”等.例句:The best kept secret for avoiding abdominal weight 〈e1〉gain〈/e1〉 due to 〈e2〉stress〈/e2〉 is the use of adaptogens.

2) 隐式因果

无因果连接词.例句:〈e1〉Water〈/e1〉 〈e2〉erosion 〈/e2〉 is the detachment and removal of soil material by water.

4.1.5 数据统计

实验共有3 000个含有因果关系的句子,按4∶1∶1的比例分为训练集、验证集、测试集.据4.1.3节所述,根据原因与结果的对应关系将句子分为一因一果、一因多果、多因一果与多因多果.由于数据规模有限(如4.3.3节的表5所示,测试集中含多因多果的句子仅有18个),为了便于实验,本文将一因多果、多因一果统称为多因多果;据4.1.4节所述,根据是否含有因果连接词将句子分为显式因果(显式连接词)、显式因果(模糊连接词)以及隐式因果.

数据统计如表2所示.由表2可知,数据中大部分为一因一果和带有显式连接词的显式因果关系,含隐式因果的句子十分稀少.

Table 2 Data Statistics

表2 数据统计

Data	ALL	One-Causality	Multi-Causality	Explicit Causality (with Explicit Connectives)	Explicit Causality (with Ambiguous Connectives)	Imexplicit Causality
All	3 000	2 352	648	1 975	838	187
Train	2 000	1 587	413	1 338	549	113
Val	500	378	122	320	140	40
Test	500	387	113	317	149	34

4.2 实验内容

4.2.1 参数设置

1) 整体实验

优化器:Adam;词向量维度:300;梯度裁剪值:5.0;学习率:0.001;迭代次数:120.

2) Bi-LSTM层

隐藏层:300;丢失率:0.3.

3) S-GAT层

加强后的特征维度 $F'=50$(原特征维度 $F=300$);多头注意力机制头的个数 $K=8$;S-GAT层数 $n'=2(n=1)$.

4.2.2 评估标准

1) 细粒度抽取准确率

以句子为单位.模型对句子进行序列标注,根据

标签序列的结果判断因果抽取是否正确.句子中所有单词的标签全部正确,则该句子的因果关系抽取正确,其中包括:

① 原因和结果抽取的单词正确;

② 因果关系方向正确;

③ 原因和结果同时抽取;

④ 在多因多果中,多个原因和多个结果同时满足上述3个条件,则抽取正确.

序列标注的准确率($accuracy$)计算为

$$accuracy=m/M, \tag{13}$$

其中,m 为句中所有单词的标签全部标注正确的句子个数,M 为句子总数.

2) 粗粒度抽取精确率(P)、召回率(R)、$F1$ 值

以标签为单位.实验是对句子中的每个单词根据标签进行三分类,即判断单词属于"原因"(C)、"结果"(E)还是"其他"(O),故实验对比不同模型的3种标签"C","E","O"的 P,R,$F1$ 值.由于句子中大部分单词的标签为"O",且实验的目的是抽取标签"C"(原因)和"E"(结果),故标签"O"的 P,R,$F1$ 值不是本文比较的重点.

4.2.3 对比模型

本文选取多种模型进行对比实验,具体包括:

1) Bi-LSTM+CRF.文献[18]提出的序列标注模型.本文将原模型中的标签修改为4.1.3节所述的因果标签,将其应用在因果关系抽取中.

2) Bi-LSTM+self-ATT.文献[20]提出的语义角色标注模型.本文将模型中的语义角色标签修改为4.1.3节所述的因果语义标签进行因果关系抽取.

3) Bi-LSTM+CRF+self-ATT.对文献[20]提出的Bi-LSTM+self-ATT模型进行改进,在注意力层后加入CRF层输出因果标签.

4) L-BL.文献[13]提出的基于语言信息的双向长短期记忆网络(linguistically informed Bi-LSTM)因果关系抽取模型.

除此之外,还包括基准模型:CRF,LSTM,LSTM+CRF,Bi-LSTM.

本文提出的因果关系抽取模型Bi-LSTM+CRF+S-GAT,根据3.2节中句法依存图生成方式的不同,有4种变形.

1) 有向图模型:Bi-LSTM+CRF+S-GAT(dir);

2) 无向图模型:Bi-LSTM+CRF+S-GAT(undir);

3) 有向图+"self-arc"模型:Bi-LSTM+CRF+S-GAT(dir+self);

4) 无向图+"self-arc"模型:Bi-LSTM+CRF+S-GAT(undir+self).

在Bi-LSTM+CRF+self-ATT模型中,设Bi-LSTM层+self-ATT层为1个块(block),堆叠 N 层,self-ATT层堆叠次数 $n'=4$.如表3所示,Bi-LSTM+CRF+self-ATT-3表示块堆叠3层,未作标记模型中的块不进行堆叠.

Table 3 Accuracy of Fine-grained Extraction

表3 细粒度抽取准确率

Model	Overall	One Causality	Multi Causality	Explicit Causality (with Explicit Connectives)	Explicit Causality (with Ambiguous Connectives)	Imexplicit Causality
CRF	0.1200	0.1214	0.1150	0.1388	0.0872	0.0882
LSTM	0.3100	0.3307	0.2389	0.3218	0.2819	0.2647
LSTM+CRF	0.3580	0.3721	0.3097	0.4101	0.2617	0.2941
Bi-LSTM	0.7060	0.7545	0.5398	0.7539	0.6711	0.4118
Bi-LSTM+CRF	0.7140	0.7468	0.6018	0.7571	0.6846	0.4412
Bi-LSTM+self-ATT	0.7200	0.7519	0.6106	0.7823	0.6443	0.4706
Bi-LSTM+CRF+self-ATT	0.7240	0.7494	0.6372	0.7760	0.6846	0.4118
L-BL	0.7320	0.7571	0.6460	0.7981	0.6577	0.4412
Bi-LSTM+CRF+self-ATT-3	0.7360	0.7726	0.6106	0.7855	0.6779	0.5294
Bi-LSTM+CRF+S-GAT(undir)	0.7320	0.7726	0.5929	0.7760	0.7047	0.4412
Bi-LSTM+CRF+S-GAT(dir+self)	0.7280	0.7571	0.6283	0.7823	0.6711	0.4706
Bi-LSTM+CRF+S-GAT(undir+self)	0.7240	0.7571	0.6106	0.7697	0.7047	0.3824
Bi-LSTM+CRF+S-GAT(dir)	0.7860	0.8243	**0.6549**	0.8076	**0.8054**	0.5000
Bi-LSTM+CRF+S-GAT(dir)-2	0.7920	0.8346	0.6460	**0.8297**	0.7718	0.5294
Bi-LSTM+CRF+S-GAT(dir)-3	**0.8000**	**0.8424**	**0.6549**	**0.8297**	0.7919	**0.5588**

Note: The numbers in bold represent the best results of the experiment.

4.3 实验结果与分析

细粒度抽取(序列标注)准确率如表3所示,粗粒度抽取("C","E","O"标签)的 P,R,$F1$ 值如表4所示:

Table 4 Value of Precision, Recall, $F1$ of Coarse-grained Extraction

表4 粗粒度抽取精确率、召回率、$F1$ 值

Model	C-P	C-R	C-$F1$	E-P	E-R	E-$F1$	O-P	O-R	O-$F1$
CRF	0.2818	0.3220	0.2883	0.4761	0.4994	0.4764	0.9267	0.9774	0.9499
LSTM	0.5884	0.6550	0.6022	0.6154	0.6910	0.6358	0.9619	0.9733	0.9665
LSTM+CRF	0.6425	0.7078	0.6587	0.6389	0.6946	0.6538	0.9675	0.9768	0.9714
Bi-LSTM	0.8545	0.8454	0.8436	0.8830	0.8907	0.8810	0.9845	**0.9928**	0.9883
Bi-LSTM+CRF	0.8527	0.8550	0.8504	0.8903	0.9001	0.8887	0.9852	0.9913	0.9879
Bi-LSTM+self-ATT	0.8702	0.8692	0.8652	0.8967	0.9128	0.8959	0.9877	0.9868	0.9870
Bi-LSTM+CRF+self-ATT	0.8667	0.8655	0.8615	0.8697	0.8765	0.8670	0.9852	0.9894	0.9870
L-BL	0.8833	0.8912	0.8822	0.8803	0.8840	0.8762	0.9859	0.9919	0.9886
Bi-LSTM+CRF+self-ATT-3	0.8853	0.8937	0.8848	0.8760	0.8658	0.8671	0.9859	0.9902	0.9879
Bi-LSTM+CRF+S-GAT(undir)	0.8821	0.8911	0.8805	0.8883	0.8953	0.8872	0.9870	0.9906	0.9885
Bi-LSTM+CRF+S-GAT(dir+self)	0.8703	0.8667	0.8649	0.8777	0.8977	0.8807	0.9861	0.9912	0.9884
Bi-LSTM+CRF+S-GAT(undir+self)	0.8560	0.8655	0.8556	0.8828	0.8857	0.8794	0.9846	0.9922	0.9881
Bi-LSTM+CRF+S-GAT(dir)	0.8975	0.9045	0.8970	0.9092	**0.9290**	0.9114	**0.9918**	0.9915	**0.9914**
Bi-LSTM+CRF+S-GAT(dir)-2	0.9050	0.9040	0.9013	0.9140	0.9067	0.9057	0.9888	0.9915	0.9900
Bi-LSTM+CRF+S-GAT(dir)-3	**0.9168**	**0.9207**	**0.9144**	**0.9182**	0.9147	**0.9125**	0.9899	0.9899	0.9898

Note: The numbers in bold represent the best results of the experiment.

4.3.1 细粒度抽取准确率

如表3所示,本文提出的 Bi-LSTM+CRF+S-GAT 模型优于其他模型,其中有向图、块堆叠3层的 Bi-LSTM+CRF+S-GAT(dir)-3 模型的准确率最高,达到了0.8,较 Bi-LSTM+CRF+self-ATT-3 序列标注模型提高了0.064,较 L-BL 因果关系抽取模型提高了0.068.后续提到的"本文提出的模型"或"Bi-LSTM+CRF+S-GAT 模型"在没有强调的情况下默认为 Bi-LSTM+CRF+S-GAT(dir)-3 模型.以图8所示的句子为例,单词"causes"在计算自身注意力时,为其在句法依存图中所有的相邻节点("flu","cold",".")分配不同的权重,使注意力更集中在单词"flu"(原因)和"cold"(结果)上,增强了因果语义的特征,提高了因果抽取的准确率.

本文提出的模型中,在块不堆叠的条件下,无向图(undir)、有向图+"self-arc"(dir+self)、无向图+"self-arc"(undir+self)这3种句法依存图模型的准确率较低,相比 Bi-LSTM+CRF+self-ATT 模型仅有轻微的提升,效果远不如有向图(dir)模型.原因和结果是同时存在且相互依赖的,是具有方向性的,无向图忽略了因果方向性的特征,导致准确率降低;单词本身不是自己的邻居,带有"self-arc"的图在更新权重时增加了干扰信息,降低了准确率.该实验结果说明 GAT 中图的选取会影响实验的结果.

在块不堆叠的情况下,改进后的 Bi-LSTM+CRF+self-ATT 模型相比文献[20]提出的 Bi-LSTM+self-ATT 模型的准确率提高了0.004,其他模型加入 CRF 层后的效果均略有改善.CRF 可以获取相邻上一词的特征,通过多种特征函数给标签打分再进行加权求和得到预测结果,是全局最优化值,故相比通过 softmax 分类器得到预测结果的准确率略有提升.对于基准模型 CRF,输入仅为词向量,以往因果关系抽取的研究中是通过人工构建特征来提高模型的准确度,故未加入任何特征工程的 CRF 模型的准确率十分低.

一因一果的准确率高于多因多果.多因多果准确率较低的原因有:

1) 存在标签具有争议无法确定的实体,句子无法给出准确的标注序列.含多因多果的句子语义较为复杂,且个人对多因多果的理解存在差异,例如句子"The slow suffocation of the cells swiftly cause unconsciousness and shock, soon followed by death.",有人认为结果为"unconsciousness"与"shock","death"是伴随行为不是结果,而有人认为"death"

是“suffocation”导致的最终结果,该句子为一因二果还是一因三果存在争议,单词“death”的标签是“E”还是“O”无法确定;在含连锁因果关系的句子中,存在既是原因又是结果的实体,例如句子“The aircraft was written off in the accident due to the severe impact caused by the KLM aircraft,and the resulting fire.”,由语义可知“impact”是“accident”的原因同时也是“aircraft”,“fire”的结果,“impact”的标签是“C”还是“E”无法确定.由于数据自身具有争议,存在标签无法确定的实体,其数据特征并不精确,故模型无法提取准确的因果特征,易导致多因多果抽取失败.

2) 因果实体数量较多.抽取多因多果时需要将句子中所有原因与结果的实体同时抽取出来,且需保证所有因果实体的中心词选取正确,难度较大.

显式因果的准确率高于隐式因果.隐式因果准确率较低的原因有:1)数据中含有隐式因果的句子较少,无法满足实验所需;2)不含因果连接词,句法结构杂乱无章,难以提取因果特征.显式因果中,显式连接词的准确率略高于模糊连接词,是因为含显式连接词的句子因果语义明显,句法结构相对规整,故准确率较高;模糊连接词的因果语义特征较为模糊,句法结构相对复杂.

4.3.2 粗粒度抽取精确率、召回率、$F1$ 值

如表 4 所示,本文提出模型的标签“C”,“E”的精确率相比 Bi-LSTM+CRF+self-ATT-3 模型分别提高了 0.031 5,0.042 2;召回率分别提高了 0.027 0,0.048 9;$F1$ 值分别提高了 0.029 6,0.045 4.

4.3.3 复杂因果对应关系分析

对于多因多果,可以细分为一因多果、多因一果与真正的多因多果(多个原因导致多个结果),其实验数据如表 5 所示:

Table 5 Data of Multi Causality

表 5 多因多果数据

Data	One-cause Multi-effect	Multi-cause One-effect	Multi-cause Multi-effect
All	297	271	80
Train	207	163	43
Val	48	55	19
Test	42	53	18

本文提出的模型 Bi-LSTM+CRF+S-GAT(dir)-3 与对比模型 Bi-LSTM+CRF+self-ATT-3、L-BL 在一因多果、多因一果与多因多果的实验数据上的准确率如表 6 所示:

Table 6 Accuracy of One-cause Multi-effect, Multi-cause One-effect, Multi-cause Multi-effect

表 6 一因多果、多因一果、多因多果的准确率

Model	One-cause Multi-effect	Multi-cause One-effect	Multi-cause Multi-effect
L-BL	0.547 6	0.698 1	0.722 2
Bi-LSTM+CRF+self-ATT-3	0.523 8	0.698 1	0.555 5
Bi-LSTM+CRF+S-GAT(dir)-3	0.547 6	0.735 8	0.666 6

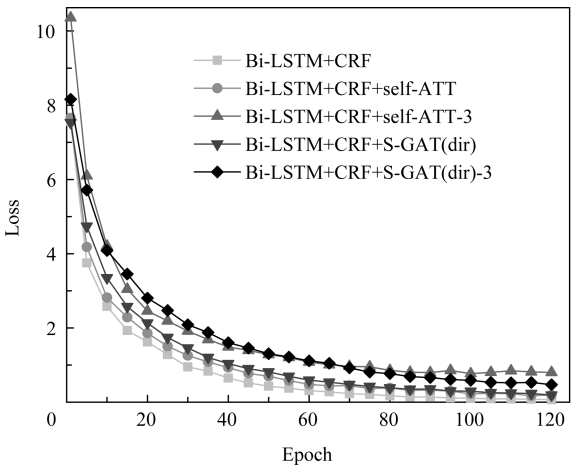
如 4.2.2 节所述,句子中所有的实体标签全部识别正确则该句子因果抽取正确,故句中每个实体标签的预测效果均会影响该句子的因果抽取结果.由于模型对于原因实体与结果实体的识别能力具有差异(C-$F1$≠E-$F1$),故句子中原因与结果实体个数的多少(因果实体个数比例=原因实体个数/结果实体个数)会影响因果抽取的准确率.由表 4 可知,L-BL,Bi-LSTM+CRF+self-ATT-3,Bi-LSTM+CRF+S-GAT(dir)-3 这 3 个模型中标签“C”(原因)的 $F1$ 值均略高于标签“E”(结果)的 $F1$ 值,说明模型对于“原因”的识别能力略优于“结果”.通常情况下,含原因($F1$ 值较高)实体个数越多的句子(如多因一果)准确率相对越高,反之含结果($F1$ 值较低)实体个数较多(如一因多果)会“拉低”该句子因果抽取正确的概率,故 3 种模型中多因一果的准确率均高于一因多果.多因多果由于因果实体个数的比例(原因实体与结果实体哪种较多)不能确定,且测试集数据量较少(仅 18 个数据),其实验结果不具有普遍代表性,故多因多果的准确率无法与一因多果和多因一果进行比较.

4.4 训练集的损失值与验证集的准确率

对比 6 种模型 Bi-LSTM+CRF, L-BL, Bi-LSTM+CRF+self-ATT, Bi-LSTM+CRF+self-ATT-3, Bi-LSTM+CRF+S-GAT(dir), Bi-LSTM+CRF+S-GAT(dir)-3 训练集的损失值(loss)与验证集的准确率,曲线如图 9,10 所示.本文选取迭代过程中验证集准确率最高的模型进行测试,结果如表 3,4 所示.

由图 9 可知,在块堆叠层数相同的情况下,本文提出模型在训练过程中的收敛速度略慢于其他模型.同一模型随着块堆叠层数的增多,损失值增大.如图 10 所示,本文提出的模型在训练初始阶段验证集准确率的上升速度较为缓慢,迭代 25 次左右时有明显上升,50 次后趋于平缓.随着迭代次数的增多,准确率逐渐高于其他模型,迭代 110 次时达到顶峰.

其他模型迭代 30 次左右时,验证集的准确率趋近于平缓,60 次迭代后趋于稳定,并缓慢上升.模型随着块堆叠层数的增加,验证集准确率的上升速度略有减缓,趋于平缓后相比块不堆叠的模型的准确率有所上升.



There is no CRF layer in the L-BL model, which is different from the calculation standard of loss value of other models and cannot be compared, so there is no L-BL model.

Fig. 9 Epoch-loss(train)

图 9 迭代次数-损失值(训练集)

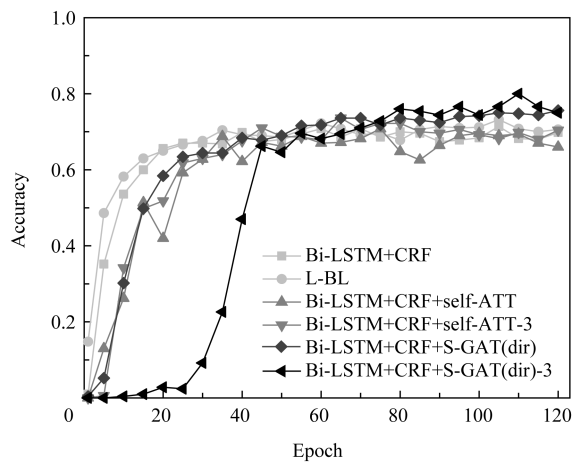


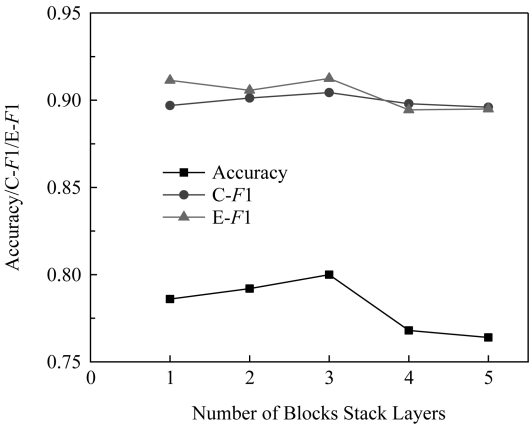
Fig. 10 Epoch-accuracy(val)

图 10 迭代次数-准确率(验证集)

在传统注意力机制的基础上,本文模型加入了图的概念,模型结构更为复杂,训练参数较多,故训练过程中的收敛速度与验证集准确率的上升速度略慢于对比模型.同一模型随着块堆叠层数的增多,网络结构的复杂程度成倍增加,故损失值增大,但提取的特征更为精准,故准确率上升.

4.5 超参数选取

本文提出的 Bi-LSTM+CRF+S-GAT(dir)模型随着块堆叠层数的增加(N 取值为 1,2,3,4,5),序列标注的准确率(细粒度抽取)与标签"C","E"的 $F1$ 值(粗粒度抽取)如图 11 所示:



C-$F1$ represents the $F1$ value of label "C",
E-$F1$ represents the $F1$ value of label "E".

Fig. 11 Number of block stack layers-accuracy/C-$F1$/E-$F1$

图 11 块堆叠层数-准确率/C-$F1$/E-$F1$

N 取值为 2,3 时,相比 $N=1$ 时模型的准确率分别提高了 0.6,1.4.$N=3$ 时,标签"C","E"的 $F1$ 值较 $N=1$ 时分别提高了 0.0174,0.0011.N 取值为 4,5 时,实验结果因过拟合下降,故模型中块的堆叠层数 $N=3$.

4.6 错误分析

模型预测错误的数据类型及其所占百分比(占比前 3)与错误例子如表 7 所示.除表 7 中的错误数据类型外,还有 3 种错误数据类型及其所占比例,分别为:原因和结果没有同时抽取占 0.05;因果关系方向错误占 0.04;连锁因果关系抽取错误占 0.04.

1) "Other".其他类型错误,因模型自身结构导致因果关系抽取失败.

2) 中心词选取错误.包括名词短语、带有"of"的短语、含有"including"等词的句子,因中心词选取不当导致预测准确率降低.如表 7 的例句中,真实数据中标签为"E"的单词是"pain",而模型预测的结果为单词"pain"和"relief",即短语"pain relief".由语义可知,相比单词"疼痛",短语"缓解疼痛"更能完整地表达因果含义.这是由于因果边界设定过于严格,中心词的选取存在歧义导致的因果抽取错误.

3) 多因多果错误.多因多果抽取正确的判定条件较为苛刻,同时识别出句子中所有原因和结果的实体才算抽取正确.如表 7 的例句所示,句子为"四因三果",预测结果因单词"burrows"的标签的识别

错误(只抽取出“三因三果”)导致整个句子的因果抽取失败.

此外,人工标注的错误与数据本身存在的争议也会导致因果抽取的准确率下降.

Table 7 Model Predicts the Wrong Data Types and their Percentages and Error Examples

表 7 模型预测错误的数据类型及其所占百分比与错误例子

Wrong Data Type	Proportion	Ture Data	Predicted Result
Other	0.41	Licenses and permits are 〈e1〉revenues〈/e1〉 from the 〈e2〉selling〈/e2〉 of vendor and dog licenses and other items.	〈e1〉Licenses〈/e1〉 and 〈e1〉permits〈/e1〉 are revenues from the 〈e2〉selling〈/e2〉 of vendor and dog licenses and other items.
Incorrect Choice of Center Word	0.25	Get neck 〈e2〉pain〈/e2〉 relief by 〈e1〉easing〈/e1〉 tension in the shoulders and upper back.	Get neck 〈e2〉pain relief〈/e2〉 by 〈e1〉easing〈/e1〉 tensionin the shoulders and upper back.
Multi-Causality Error	0.20	The mite's 〈e1〉burrows〈/e1〉, fecal 〈e1〉matter〈/e1〉, 〈e1〉proteins〈/e1〉 and 〈e1〉eggs〈/e1〉 cause 〈e2〉itching〈/e2〉, 〈e2〉rashes〈/e2〉 and 〈e2〉sensitivity〈/e2〉.	The mite's burrows, fecal 〈e1〉matter〈/e1〉, 〈e1〉proteins〈/e1〉 and 〈e1〉eggs〈/e1〉 cause 〈e2〉itching〈/e2〉, 〈e2〉rashes〈/e2〉 and 〈e2〉sensitivity〈/e2〉.

5 总 结

因果关系是一种重要的关系类型,因果关系抽取是文本挖掘中的一项基本任务.与传统的文本分类或关系抽取的分类方法不同,序列标注能够抽取出句子中的因果实体,且识别出因果关系方向,做到真正的因果“抽取”.结合传统的序列标注模型与注意力机制,本文提出了 Bi-LSTM+CRF+S-GAT 因果抽取模型.该模型将 GAT 应用到 NLP 中,拓展句法依存树到句法依存图并引入了 S-GAT.将传统注意力机制中的线性数据转化为图形数据,原本互相独立的词通过句法依存图产生依赖关系.句中的词在计算自身注意力时,为每个相邻节点分配不同的权重,使注意力更集中在表示“原因”与“结果”含义的词上,加强了因果语义特征,达到了因果抽取的目的.本文提出的模型较现有 Bi-LSTM+CRF+self-ATT 模型的准确率提高了 0.064.

由于语义的复杂性和标注的歧义性,本文提出的因果关系抽取方法还存在一些缺点与不足:1)标注方法的缺陷.因果边界的选取过于严格,英文中有时短语更能完整地表达因果含义,连锁因果关系无法给出准确的标注序列.2)抽取方法的缺陷.序列标注的方法无法对原因或结果是子句(无法提取因果中心词)的句子进行因果关系抽取.3)实验数据的缺陷.因果关系抽取可进行公开测评的数据集较少,本文采用人工扩展 SemEval 数据集后的实验数据依旧不多,人工标注速度较慢且存在错误.实验数据中不涉及跨句、跨段的因果关系,且含隐式因果的数据量较少,无法深入探究隐式因果关系.

此外,GAT 中图的选取会影响因果抽取的效果,除本文采用的句法依存图外,也可尝试语义依存图或构建其他类型的图进行研究.进一步拓展因果抽取范围,如深入研究隐式因果关系,探究跨句、跨段等长文本的因果关系抽取.除因果关系抽取外,本文提出的模型也可尝试应用在其他序列标注的任务中.这些均是我们以后需要改进和进一步探究的问题.

参 考 文 献

[1] Radinsky K, Davidovich S, Markovitch S. Learning causality for news events prediction [C] //Proc of the 21st Int Conf on World Wide Web. New York: ACM, 2012: 909-918

[2] Girju R. Automatic detection of causal relations for question answering [C] //Proc of the ACL 2003 Workshop on Multilingual Summarization and Question Answering-Volume 12. Stroudsburg, PA: ACL, 2003: 76-83

[3] Hashimoto C, Torisawa K, Kloetzer J, et al. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features [C] //Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2014: 987-997

[4] Velickovic P, Cucurull G, Casanova A, et al. Graph attention networks [J]. arXiv preprint arXiv:1710.10903, 2017

[5] Garcia D. COATIS, an NLP system to locate expressions of actions connected by causality links [C] //Proc of the Int Conf on Knowledge Engineering and Knowledge Management. Berlin: Springer, 1997: 347-352

[6] Khoo C S G, Kornfilt J, Oddy R N, et al. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing [J]. Literary and Linguistic Computing, 1998, 13(4): 177-186

[7] de Silva T N, Xiao Zhibo, Zhao Rui, et al. Causal relation identification using convolutional neural networks and knowledge based features [J]. International Journal of Computer and Systems Engineering, 2017, 11(6): 697-702

[8] Hidey C, McKeown K. Identifying causal relations using parallel Wikipedia articles [C] //Proc of the 54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2016: 1424-1433

[9] Zhao Sendong , Liu Ting , Zhao Sicheng , et al. Event causality extraction based on connectives analysis [J]. Neurocomputing, 2016, 173: 1943-1950

[10] Feng Chong , Kang Liqi , Shi Ge, et al. Causality extraction with GAN [J]. Acta Automatica Sinica, 2018, 44(5): 811-818 (in Chinese)
(冯冲, 康丽琪, 石戈, 等. 融合对抗学习的因果关系抽取[J]. 自动化学报, 2018, 44(5): 811-818)

[11] Kruengkrai C, Torisawa K, Hashimoto C, et al. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 3466-3473

[12] Fu Jianfeng , Liu Zongtian , Li Wei, et al. Event causal relation extraction based on cascaded conditional random fields [J]. Pattern Recognition and Artificial Intelligence, 2011, 24(4):567-573 (in Chinese)
(付剑锋, 刘宗田, 刘炜, 等. 基于层叠条件随机场的事件因果关系抽取[J]. 模式识别与人工智能, 2011, 24(4): 567-573)

[13] Dasgupta T, Saha R, Dey L, et al. Automatic extraction of causal relations from text using linguistically informed deep neural networks [C] //Proc of the 19th Annual SIGdial Meeting on Discourse and Dialogue. Stroudsburg, PA: ACL, 2018: 306-316

[14] Qiu Jiangnan, Xu Liwei, Zhai Jie, et al. Extracting causal relations from emergency cases based on conditional random fields [J]. Procedia Computer Science, 2017, 112: 1623-1632

[15] Feng Xuewei, Wang Dongxia, Huang Minhuan, et al. A mining approach for causal knowledge in alert correlating based on the Markov property [J]. Journal of Computer Research and Development, 2014, 51(11): 2493-2504 (in Chinese)
(冯学伟, 王东霞, 黄敏桓, 等. 一种基于马尔可夫性质的因果知识挖掘方法[J]. 计算机研究与发展, 2014, 51(11): 2493-2504)

[16] Luo Zhiyi, Sha Yuchen, Zhu K Q, et al . Commonsense causal reasoning between short texts [C/OL] //Proc of the 15th Int Conf on Principles of Knowledge Representation and Reasoning. Menlo Park, CA: AAAI, 2016 [2019-12-21]. https://www.aaai.org/ocs/index.php/KR/KR16/paper/viewPaper/12818

[17] Lafferty J, McCallum A, Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C] //Proc of the 18th Int Conf on Machine Learning. New York: ACM, 2001: 282-289

[18] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional LSTM-CRF models for sequence tagging [J]. arXiv preprint arXiv:1508.01991, 2015

[19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C/OL] //Proc of the 31st Conf on Neural Information Processing Systems. 2017: 5998-6008 [2019-07-20]. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[20] Tan Zhixing, Wang Mingxuan, Xie Jun, et al. Deep semantic role labeling with self-attention [C/OL] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018 [2019-12-21]. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16725

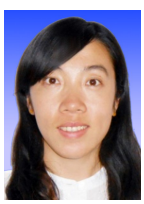
[21] Kipf T N, Welling M. Semi-supervised classification with graph convolutional networks [J]. arXiv preprint arXiv:1609.02907, 2016

[22] Ittoo A, Bouma G. Extracting explicit and implicit causal relations from sparse, domain-specific texts [C] //Proc of the Int Conf on Application of Natural Language to Information Systems. Berlin: Springer, 2011: 52-63

Xu Jinghang, born in 1995. Master candidate at Jilin University. Her main research interests include natural language processing, machine learning.

Zuo Wanli, born in 1957. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include Web mining, natural language processing, machine learning, deep learning and Web search engines.

Liang Shining, born in 1994. Master candidate at Jilin University. His main research interests include natural language processing and deep learning.



Wang Ying, born in 1981. PhD, associate professor. Senior member of CCF. Her main research interests include data mining, machine learning, social computing and search engine. (wangying2010@jlu.edu.cn)