

融合关系标签和位置信息的中文医疗文本因果关系抽取方法研究^{*}

张维宁¹ 申喜凤² 李美婷³ 高东平¹

(¹ 中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100020

²首都医科大学附属北京潞河医院 北京 101149 ³黑龙江省高级人民法院 哈尔滨 150090)

〔摘要〕 目的/意义 利用因果关系词相对位置辅助深度学习模型,提高因果关系预测能力,挖掘医疗文本增益信息。方法/过程 将医疗文本因果关系词相对位置信息表示为关系特征层嵌入预训练语言模型,融合基线模型进行实体识别及关系抽取。结果/结论 嵌入关系特征层的模型 *F1* 值较基线模型 BERT - BiLSTM - CRF 和 CasRel 分别提升 2.92 个百分点和 6.41 个百分点,因果关系预测能力较好。

〔关键词〕 自然语言处理;因果关系抽取;预训练模型;BERT;医疗文本

〔中图分类号〕 R - 058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.01.004

Study on the Method of Causality Extraction from Chinese Medical Texts by Integrating Relational Label and Location Information

ZHANG Weining¹, SHEN Xifeng², LI Meiting³, GAO Dongping¹

¹Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100020, China; ²Beijing Luhe Hospital, Capital Medical University, Beijing 101149, China; ³The High People's Court of Heilongjiang Province, Harbin 150090, China

〔Abstract〕 **Purpose/Significance** The relative positions of causality words are utilized to assist deep learning models to improve causality prediction and mine medical text gain information. **Method/Process** The relative position information of causality words in medical texts is represented as a relational feature layer embedded in a pre-trained language model, and the baseline model is integrated for entity recognition and relationship extraction. **Result/Conclusion** The *F1* value of the model embedded in the relational feature layer is improved by 2.92 percentage points and 6.41 percentage points compared with the baseline models BERT - BiLSTM - CRF and CasRel, respectively, with better causal prediction capacity.

〔Keywords〕 natural language processing; causality extraction; pre-training model; BERT; medical text

1 引言

因果关系是指一个事件(即“原因”)和另一事件(即“结果”)之间的作用关系,表示客观事件之间的一种普遍联系。随着线上问诊等医疗信息化产业的发展,越来越丰富的医疗问答信息和知识

〔修回日期〕 2023 - 10 - 27

〔作者简介〕 张维宁,硕士研究生,发表论文 1 篇;通信作者:高东平,研究员,硕士生导师。

〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目(项目编号:2020AAA0104905)。

类文本可以直接通过互联网获取,其中存在大量因果关系解释,对患者缓解病症有帮助,应用于医疗搜索和诊断业务具有巨大的潜力和价值,从中可以挖掘抽取医疗文本之间存在的因果逻辑关系,构建关系解释网络及医疗知识图谱,提升对医疗结果的逻辑性和可解释性的判断能力^[1]。

因果关系可分为显式因果关系和隐式因果关系,其主要区别在于是否有关系特征词对关系进行提示。因果关系特征词用于提示句子成分之间存在的显式因果关系,可分为显性关系特征词与模糊关系特征词。显性关系特征词在大多数语境下均具有明显的关系含义。模糊关系特征词只有在特定语境,结合上下文及领域专业性进行推断,才可以成为关系特征词。隐式因果关系的关系成分之间不通过关系特征词进行链接,要结合上下文进行理解和推断。

现有研究常采用深度学习方法进行因果关系抽取,语义规则也逐渐被引入关系抽取模型,用以提高关系抽取的准确率及效率。目前,已有学者^[2]基于长短期记忆网络(long short-term memory, LSTM)模型实现因果关系抽取;也有研究^[3]基于多任务思想提出二元关系抽取和一元功能识别共同决策的联合学习模式,用于挖掘生物医学实体间的因果关系;此外,多特征融合方式也被证明可以提升实体关系抽取效果^[4]。在医疗文本领域,有研究^[5]从语义谓词优化的角度对疾病因果关系抽取方法进行改进,实现了从大规模生物医学文献中抽取疾病因果关系;也有研究^[6]应用中文医疗因果关系模型从压力性损伤、心血管疾病、宫颈癌 3 个健康领域进行因果关系图构建。

借鉴以上思路,本研究通过构建医疗文本因果关系特征词库,定位因果关系特征词,将相对关系特征词位置信息引入预训练语言模型双向编码器表征(bidirectional encoder representations from transformers, BERT),结合双向长短期记忆模型(bi-directional long-term memory, BiLSTM)和条件随机场(conditional random field, CRF),同时实现对显隐式因果关系的识别,为互联网医疗文本挖掘提供新的思路和方向。

2 概念限定

2.1 医学概念片段概念限定

本研究针对互联网医疗文本中的医学概念片段进行实体识别及因果关系抽取。其中,医学概念片段是指以医学实体、临床发现和疾病症状为中心的作为独立语义单元的连续字符片段,可以是单个词语、短语,也可以是能够独立成句、结构完整的一段文字。互联网医疗文本非官方书面信息,具有口语化、实体长度较长且变化范围大、边界模糊、分词困难、难以清晰识别等特点,这为医疗文本的实体识别和因果关系判断带来一定挑战。

2.2 因果关系概念限定

本研究将医疗文本因果关系限定为某种原因直接导致某种结果,包括疾病和临床症状之间的关系,同样可分为以显性和模糊因果关系特征词充当关系提示词的显式因果关系和需要通过理解推断得到的隐式因果关系两类。

例如,在中文医疗文本“钙元素缺乏会导致神经的兴奋性增高”中,“导致”可作为“钙元素缺乏”与“神经的兴奋性增高”之间的显性因果关系特征词,提示二者之间存在的显式因果关系;又如,在“慢性胰腺炎的主要症状是腹痛”中,“症状”作为“慢性胰腺炎”与“腹痛”之间的模糊因果关系特征词出现,提示二者之间存在的特定关系及关系类型,可理解为“慢性胰腺炎”会导致“腹痛”。利用这两类因果关系特征词构建因果关系特征词库,有助于定位显式因果关系位置,辅助因果关系识别。

而在中文医疗文本“经腹腔镜的宫颈肌瘤手术,术中出血会比较多”中,通过语义理解可发现,其中存在隐式因果关系,即“经腹腔镜的宫颈肌瘤手术”会导致“术中出血”。此处的因果关系没有因果关系特征词作为提示,是通过语义推理得到的。此类隐式因果关系在中文医疗文本语料中较少,且深度学习技术强大的表征学习能力使其能够有效捕捉文本中的隐式因果关系,因此,本研究拟采用深度学习模型对隐式因果关系进行识别判断。

3 医疗文本因果关系特征词库构建

针对医学概念片段的复杂性及显式因果关系以因果关系特征词作为指示特征的特点，利用互联网中文医疗文本语料，对其中存在的因果关系进行分析，结合医疗文本的领域特性及专业性，对其中用以提示显式因果关系的显性和模糊因果关系特征词

进行分析归纳，构建医疗文本因果关系特征词库，定位因果关系特征词位置，以相对因果关系特征词的距离信息辅助医学概念片段分割，进而实现利用深度学习模型对显隐式因果关系的共同识别。

本词库包含汉语中可普遍提示因果关系的特征词汇，以及适用于临床发现、疾病症状关系的医疗文本因果关系特征词，充分体现医疗文本特点，以提高识别准确度，见表 1。

表 1 因果关系特征词（部分）

关系类别	关系特征词	例句
显式因果关系	显性因果特征词	引起 “长期抽搐发作” 时容易引起 “脑部缺血、缺氧性改变”
		导致 “钙元素缺乏” 会导致 “神经的兴奋性增高”，尤其是晚上症状明显
		由于 “浆细胞恶变” 是由于 “长期慢性的炎症刺激”
	模糊因果特征词	症状 “慢性胰腺炎” 的主要症状是 “腹痛”
	特征	“皮肤黏蛋白沉积症” 是一组以 “皮肤内黏蛋白弥漫性或局限性沉积” 为特征的疾病

4 融合关系标签和位置信息的中文医疗文本因果关系抽取方法

4.1 BERT 预训练语言模型

基于变换器的 BERT 模型^[7]由多层 Transformer 编码模块堆叠而成，通过获取输入文本的动态词向量捕获文本的深层次语义信息及更加丰富的语义表达，得到的向量矩阵由字符表征（token embeddings）、句子表征（segment embeddings）和位置表征（position embeddings）构成^[8]。将长度为 n 的文本序列表示为 $X = (x_1, x_2, \cdots, x_n)$ ，在文本序列首尾分别添加

“[CLS]” 和 “[SEP]” 标志，分别对应输入文本中第一个字符和最后一个字符的字符向量，则 BERT 输出的文本序列向量可以表示为：

$$V = \text{BERT}([\text{CLS}], x_1, x_2, \cdots, x_n, [\text{SEP}]) = (e_1, e_2, \cdots, e_n)$$

(1)

其中， x_i 表示输入序列的第 i 个文字， V 为输入序列 X 的向量化表示。

利用医学概念片段边界与因果关系特征词位置的相关性，抽取得到一个相对关系特征词位置的向量，用于表示每一个字符（token）与最近一个关系特征词的相对距离，并将此向量定义为关系表征，与 BERT 模型中已有的 3 类表征信息进行拼接，得到每个 token 的最终向量，见图 1。

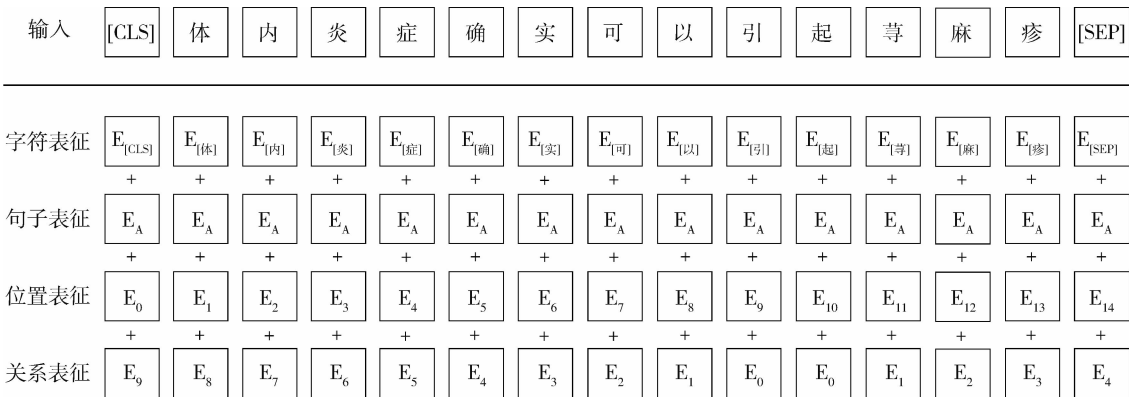


图 1 本研究中 BERT 预训练模型的嵌入表示

4.2 BiLSTM 模型

LSTM^[9]是对传统循环神经网络（recurrent neural networks, RNN）^[10]的改进，可获得长跨度单词的特征。选择 BiLSTM 模型^[11]获取上下文中文本的特征，输出的特征向量由从前向后的 LSTM 和从后向前的 LSTM 输出拼接而成，包括前向和后向的文本特征^[12]。对于从 BERT 输出的序列向量，使用 LSTM 单元进行处理，计算过程如下：

$$\vec{h}_i = \overrightarrow{LSTM}(e_i, \vec{h}_{i-1})$$

$$\overleftarrow{h}_i = \overleftarrow{LSTM}(e_i, \overleftarrow{h}_{i+1})$$

其中，输入向量为 e_i ，前向 LSTM 的隐藏状态为 $\vec{h}_i$ ，后向 LSTM 的隐藏状态为 $\overleftarrow{h}_i$ ，将二者拼接得到 BiLSTM 输出的表示为 $h_i = [\vec{h}_i, \overleftarrow{h}_i]$ 。

4.3 CRF 层

CRF 是在给定一组随机输入变量的条件下，输出一组随机变量的条件概率分布模型^[13]。模型将每个字所对应全部标签的发射概率矩阵和初始化的转

移矩阵作为原始模型的训练参数，采用最大似然函数进行训练，通过学习特征到标注结果的映射，得到特征到任意标签的条件概率 $P(Y|X)$ ，其中， Y 为输出变量， X 为输入变量。例如输入 x 计算得到条件概率最大的输出序列 y ，确定标签序列，获得最终关系预测结果。

4.4 BERT + BiLSTM + CRF 模型

本研究提出的模型应用于医疗概念片段的实体识别及因果关系抽取，模型结构主要由 3 部分组成，分别为嵌入关系特征层的预训练层、BiLSTM 神经网络层及 CRF 分类层，见图 2。首先，使用预训练语言模型对语料数据进行上下文信息编码嵌入，获得包含字符向量、句子向量、位置向量及相对关系词位置向量的最终向量，提升向量表示能力；其次，使用 BiLSTM 模型结构获取结合上下文信息的长期依赖特征；最后，利用 CRF 层完成因果关系抽取，输出最终关系分类结果。

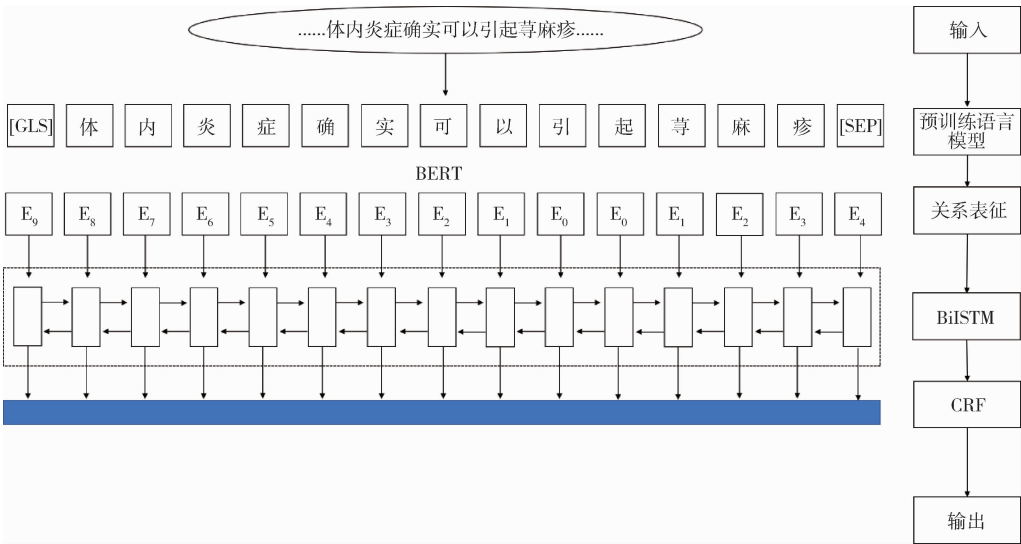


图2 模型结构

5 实验设计

5.1 数据准备

本实验数据集 CMedCausal^[14] 来源于阿里云天池平台，由阿里巴巴夸克医疗事业部和阿里云天池

联合提供，包括 800 段标注语料作为训练数据，199 段标注语料作为测试数据，见表 2。

表 2 数据集描述

数据集	最大字符长度	平均字符长度	最小字符长度	因果关系数
训练集	520	267.7	29	5 630
测试集	544	264.6	129	1 426

5.2 实验环境及参数设置

具体实验环境配置，见表 3。采用 12 层 Transformer 模型，隐藏层维度设置为 768，最大序列长度设置为 500，学习速率为 0.000 3，批大小为 16，训练次数为 500。

表 3 实验环境配置

实验环境	配置
操作系统	Windows 10 专业版
CPU	Intel (R) Xeon (R) Gold 6234 CPU @ 3.30GHz
内存	192GB
编程语言	Python 3.6.13
开发环境	PyCharm2023.1.3
GPU	Tesla T4 - 15GB

5.3 评价指标

以因果关系三元组的准确率 accuracy、召回率 recall 及 F1 作为评估指标：

$$\text{accuracy} = \frac{\text{correct}(n)}{\text{proposed}(n)}$$

(4)

$$\text{recall} = \frac{\text{correct}(n)}{\text{proper}(n)}$$

(5)

$$F1 = \frac{2\text{accuracy} \times \text{recall}}{\text{accuracy} + \text{recall}}$$

(6)

其中，correct (n) 表示模型正确识别的因果关系个数；proposed (n) 表示模型识别的所有因果关系个数；proper (n) 表示数据集中真实的因果关系个数。

6 实验结果分析

6.1 对比模型

为验证本研究中嵌入关系特征层的 BERT + BiLSTM + CRF 模型抽取因果关系的性能，进行对比实验，其中包括 3 个基线模型。一是 BiLSTM + CRF^[11]，对输入字符序列的上下文语义进行学习，通过 CRF 层输出全局最优的标记序列，表现较好。二是 BERT + CRF^[15]，结合 BERT 预训练语言模型学习字符序列的状态特征，并将得到的序列状态分数输入 CRF 层，由其对序列状态转移做出约束优化^[16]。三是 BERT + BiLSTM + CRF^[17]，利用 BERT

预训练语言模型进行字符向量训练工作，后接 BiLSTM 层和 CRF 层，导入输出的文本字符嵌入向量。还包括 1 个主流三元组抽取模型 CasRel^[18]，基于参数共享的联合实体关系抽取方法，将该模型中的 BERT 预训练模型进行嵌入关系特征层替换，对模型结果进行对比分析。

6.2 实验结果

本研究提出的嵌入关系特征层的 BERT + BiLSTM + CRF 模型在中文医疗因果关系抽取数据集 CMedCausal 上有较好的识别能力，准确率达 65.36%，召回率达 68.47%，F1 值达 66.88%，较未嵌入关系特征层的 BERT + BiLSTM + CRF 模型有一定提升（F1 值提高 2.92 个百分点），且实验结果均优于其他对比模型。将嵌入关系特征层的 BERT 预训练模型移植到关系三元组抽取模型 CasRel 上同样表现出较好识别能力，F1 值较基准模型提升 6.41 个百分点。本研究方法的指标值较其他方法更高主要与以下原因有关：一是 BERT + BiLSTM + CRF 更适用于因果关系抽取任务；二是引入的 BERT 预训练语言模型具有更强大的文本表示能力；三是嵌入相对关系特征词位置信息层，突出因果关系的语义特征及关系边界特征，对实体边界具有更清晰的识别能力，见表 4。

表 4 本研究模型与其他模型性能对比情况

模型	准确率	召回率	F1 (%)
BiLSTM + CRF	30.808	27.318	28.959
BERT + CRF	63.816	63.208	63.510
BERT + BiLSTM + CRF	62.931	65.013	63.955
BERT (ours) + BiLSTM + CRF	65.359	68.471	66.879
CasRel	47.400	39.522	43.122
CasRel (ours)	50.995	48.156	49.535

7 结语

本研究设计了一种融合关系标签和位置信息的关系抽取方法，通过字符嵌入融合相对关系词位置信息的方法融入了更多语义信息，有助于减少分词

错误造成的误差，提高中文医疗因果关系抽取精度。本研究提出的方法能够较好地解决医学概念片段在实体识别及关系抽取中存在的复杂性问题，相比于其他基线模型在性能上有一定程度提升。对于部分因果关系存在的缺少句子固定组成成分等问题，有待通过实体补齐等方式进行进一步研究及深入探索。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 李子昊, 陈漠沙, 马镇新, 等. 中文医疗因果关系抽取数据集 CMedCausal [J]. 医学信息学杂志, 2022, 43 (12): 23 – 27, 31.

2 DASGUPTA T, SAHA R, DEY L, et al. Automatic extraction of causal relations from text using linguistically informed deep neural networks [C]. Rome: The 19th Annual SIGdial Meeting on Discourse and Dialogue, 2018.

3 刘苏文, 邵一帆, 钱龙华. 基于联合学习的生物医学因果关系抽取 [J]. 中文信息学报, 2020, 34 (4): 60 – 68.

4 景慎旗, 赵又霖. 基于医学领域知识和远程监督的医学实体关系抽取研究 [J]. 数据分析与知识发现, 2022, 6 (6): 105 – 114.

5 董鹏, 李晓瑛, 李军莲, 等. 基于语义谓词优化的疾病因果关系发现 [J]. 数字图书馆论坛, 2022 (11): 19 – 25.

6 郭思伊. 面向健康领域的因果关系图构建方法研究 [D]. 合肥: 合肥工业大学, 2023.

7 DEVLIN J, CHANG M W, LEE K, et al. Bert: pre – training of deep bidirectional transformers for language understanding [C]. Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

8 郑巧夺, 吴贞东, 邹俊颖. 基于双层 CNN – BiGRU –

CRF 的事件因果关系抽取 [J]. 计算机工程, 2021, 47 (5): 58 – 64, 72.

9 GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: continual prediction with LSTM [J]. Neural computation, 2000, 12 (10): 2451 – 2471.

10 ELMAN J L. Finding structure in time [J]. Cognitive science, 1990, 14 (2): 179 – 211.

11 HUANG Z, XU W, YU K. Bidirectional LSTM – CRF models for sequence tagging [EB/OL]. [2023 – 07 – 17]. <https://arxiv.org/abs/1508.01991>.

12 郑余祥, 左祥麟, 左万利, 等. 基于时间关系的 Bi – LSTM + GCN 因果关系抽取 [J]. 吉林大学学报 (理学版), 2021, 59 (3): 643 – 648.

13 姜博, 左万利, 王英. 基于 BERT 的因果关系抽取 [J]. 吉林大学学报 (理学版), 2021, 59 (6): 1439 – 1444.

14 阿里云天池. 中文医疗信息处理评测基准 CBLUE [DB/OL]. [2023 – 07 – 17]. <https://tianchi.aliyun.com/dataset/95414>.

15 CAI Q. Research on Chinese naming recognition model based on BERT embedding [C]. Beijing: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS), 2019.

16 何涛, 陈剑, 闻英友. 基于 BERT – CRF 模型电子病历实体识别研究 [J]. 计算机与数字工程, 2022, 50 (3): 639 – 643.

17 JIANG S, ZHAO S, HOU K, et al. A BERT – BiLSTM – CRF model for Chinese electronic medical records named entity recognition [C]. Xiangtan: 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA), 2019.

18 WEI Z, SU J, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction [C]. Seattle: The 58th Annual Meeting of the Association for Computational Linguistics, 2020.

欢迎订阅

欢迎赐稿