基于预训练模型的药物不良事件抽取方法研究*

袁 驰[1] 李计巧[1] 王正瑶[1] 王怀玉[2]

([1] 河海大学计算机与软件学院 南京 211100
[2] 北京中医药大学国家中医体质与治未病研究院 北京 100029)

〔摘要〕 **目的/意义** 研究医学文本药物不良事件数据抽取方法，为临床用药风险管理和科学决策提供支持。**方法/过程** 基于预训练模型，结合实体识别和关系抽取两个子任务的关联性，设计面向药物不良事件监测的实体关系联合抽取方法。**结果/结论** 在公开药物不良事件抽取数据集上的实验表明，该方法优于已有方法，能够有效地从医学文本中提取药物不良事件信息及其关系，为药物不良事件的发现与监测提供有力手段。

〔关键词〕 药物不良事件；实体关系抽取；预训练模型；自然语言处理；医学
〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.02.007

A Pre-trained Language Model-based Method for Adverse Drug Events Extraction

YUAN Chi[1], *LI Jiqiao*[1], *WANG Zhengyao*[1], *WANG Huaiyu*[2]

[1] *School of Computer Science and Software Engineering, Hohai University, Nanjing 211100, China*; [2] *National Institute of Traditional Chinese Medicine Constitution and Preventive Treatment of Diseases, Beijing University of Chinese Medicine, Beijing 100029, China*

〔**Abstract**〕 **Purpose/Significance** To study the extraction method of adverse drug event (ADE) data from medical texts, and to provide support for clinical drug risk management and scientific decision-making. **Method/Process** Based on pre-trained model, by combining the correlation between the two subtasks of entity recognition and relation extraction, a entity relation joint extraction method for ADE monitoring is designed. **Result/Conclusion** Experiments on the published ADE extraction dataset show that the proposed method is superior to existing methods and can effectively extract ADE information and its relation from medical texts, providing a powerful means for the discovery and monitoring of ADE.

〔**Keywords**〕 adverse drug event; entity relation extraction; pre-trained model; natural language processing; medicine

1 引言

药物不良事件（adverse drug event，ADE）是指患者在应用药物时出现的不良临床事件，可能会导致住院、残疾甚至死亡[1]。尽管在临床试验阶段，药物研发者试图发现和减少药物使用过程中可能出现的各类不良反应，但在药物上市后仍难免有新的不良反应

〔修回日期〕 2023-08-05
〔作者简介〕 袁驰，博士，讲师，发表论文 20 余篇。
〔基金项目〕 国家自然科学基金项目（项目编号：62302151）；中央高校基本科研业务费（项目编号：B220202076，B220201032）。

事件发生[2]。统计数据显示，ADE 每年导致超过 350 万次内科就诊以及 100 万次急诊就诊[3]。

ADE 抽取作为医学信息抽取的重要任务，一直以来受到广泛关注。从最早的 ADE 数据集[4]到 2018 年 n2c2 的 ADE 评测任务[5]，丰富的 ADE 数据集为抽取方法的研究提供了有效支撑。在众多数据集上，不少研究者积极探索各类方法[6-9]。如 Li F 等[6]提出一种基于卷积神经网络的联合抽取模型，在 ADE 数据集上实验表明其优于流水线方法。实体关系联合抽取在于充分利用两个子任务的特性联合训练，避免了流水线方法中的错误累积，受到不少研究者的关注[10-14]。近年来预训练模型的引入为此研究提供了新的解决思路[15]。Giorgi J 等[16]基于预训练模型和联合抽取模型在多个公开数据集上取得了不错的效果。

2 研究方法

2.1 总体框架

本文设计一种基于预训练模型的实体关系联合抽取方法，见图 1。第 1 步：输入序列首先经过预训练语言模型，得到最终隐藏层的向量表示。第 2 步：经过命名实体识别任务模块对每个词（token）分类，输出对应的 token 实体标签，完成实体识别任务。第 3 步：根据实体识别结果确定句子中实体的边界位置，通过预训练语言模型获得对应 token 的向量表示，经过关系抽取任务模块，获取实体间的关系类别。

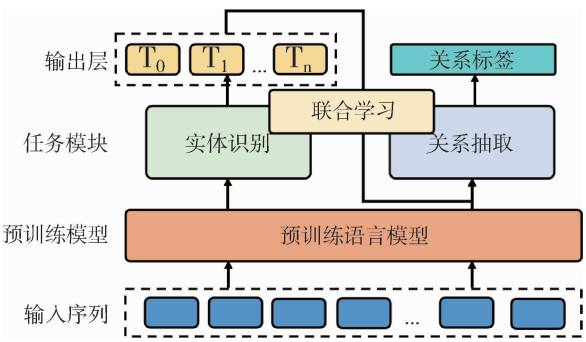

图 1 基于预训练模型的实体关系联合抽取框架

2.2 实体识别模块

针对每个句子的输入序列 S，假设其由多个 token 组成，$w_0, w_1, w_2, \ldots, w_n$，先将输入的单词序列转换为其对应的词级别的上下文嵌入表示。实体识别模块，见图 2。在基于预训练语言模型的编码器中，先通过基于 WordPiece 字典的方法将原始的单词输入序列切分化，并且在序列的首端和尾端连接［CLS］和［SEP］标志，生成［CLS］，$\widetilde{w_0}$，$\widetilde{w_1}$，$\widetilde{w_2}, \ldots, \widetilde{w_n}$，［SEP］，输入基于 Transformer 的模块进行编码，计算方式如下：

$$h_i = \text{TransformerBlock}(\widetilde{w_i}) \tag{1}$$

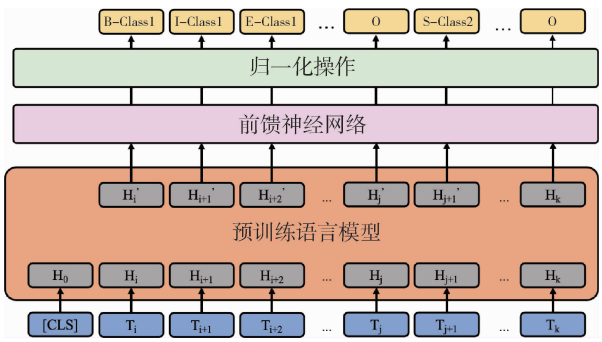

图 2 实体识别模块

实体识别模块本质上是对输入序列 S 的每个 token 分类，从而得到待识别实体和非实体之间的边界。为了充分利用实体间的上下文关系，通过将编码后 token 的上下文表示 h_i 输入一个前馈神经网络，经过一个归一化操作 softmax 层，得到每个 token 的所属标签，计算方式如下：

$$P_i = \text{softmax}(\text{FFNN}(h_i)) \tag{2}$$

本文采用预训练 BERT 模型，在编码过程中使用了基于 WordPiece 字典的切分化方法，可能出现输入序列中的单个单词被切分成多个 token 的情况。针对该问题，取首个 token 的实体标签代指整个单词的标签，避免出现同一个单词中部分属于某一实体，而剩余部分属于另一个实体的情况。在实体识别模块的训练中，采用基于交叉熵的损失函数，计算方式如下：

$$L_{\text{ner}} = -\sum_{i=1}^{n} [y^i \log P_i + (1 - y^i) \log(1 - P_i)] \tag{3}$$

2.3 关系抽取模块

2.3.1 实体关系的编码 在关系抽取模块中，受 R－BERT 方法的启发，以原数据句子序列作为输入的同时，将命名实体识别结果同时传入，作为判定实体边界的依据，见图 3。对每个输入的句子 s，为了提取其中每个实体的表示，在实体识别结果中，选取实体 1 和实体 2 中的末尾 token 作为对应实体的向量表示，再通过激活函数激活，得到实体 1 和实体 2 的编码结果，计算方式如下：

$$H'_{e1} = W_1(\tanh(H_k)) + b_1 \quad (4)$$

$$H'_{e2} = W_2(\tanh(H_v)) + b_2 \quad (5)$$

为了获得输入序列的整体表示，与 BERT 预训练模型相对应，获取每个句子序列中的首个 token，即［CLS］token 在最后一个隐藏层的结果，作为整个序列特征的表示，即图 3 中的 H_0，经过公式（6）中的激活函数激活后得到 H'_0，用作后续处理中代表整个序列的特征表示。本文采用的序列表示方法，不依赖人工设置特征表示，既不需要通过句法分析或者词法分析的结果设计特征或者核函数，也不需要设计具体复杂的深度神经网络，而 word embedding[17]、Character Embedding[18]则要通过深度学习方法进行特征表示。

$$H'_0 = W_0(\tanh(H_0)) + b_0 \quad (6)$$

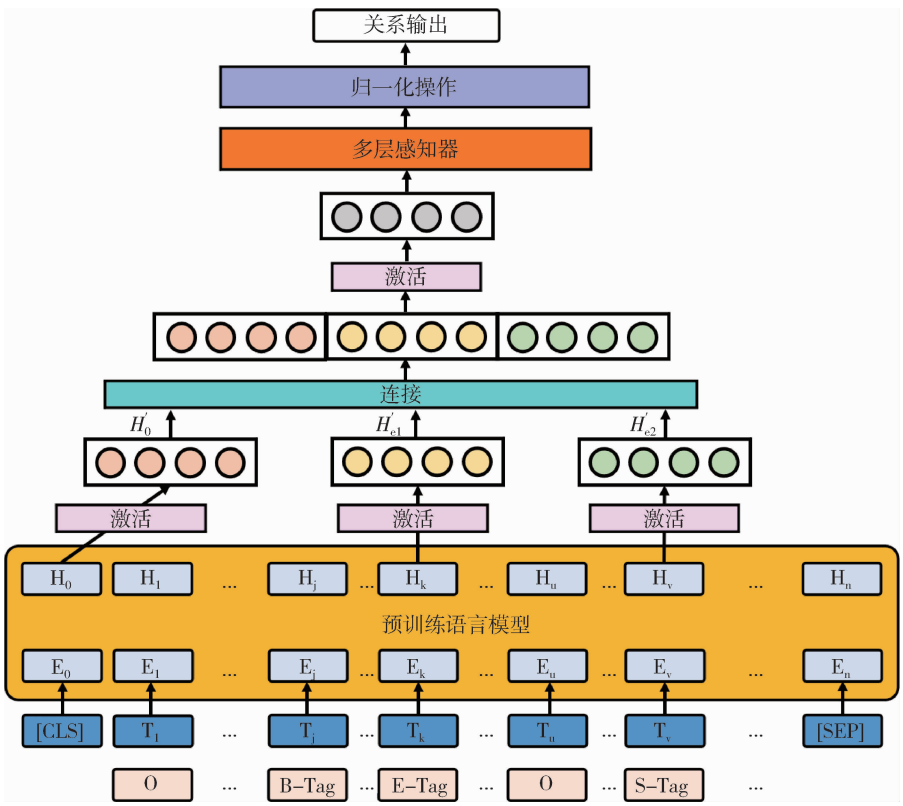

图 3 关系抽取模块

关系抽取可以转换为机器学习方法中的分类任务。在获得成对实体的表示、序列的表示后，通过对 3 个向量集联操作，获得最终用于关系分类的特征表示，计算方式如下：

$$H_{rel} = W_3(\text{concat}(H'_0, H'_{e1}, H'_{e2})) + b_3 \quad (7)$$

2.3.2 实体间关系的分类 在获得关系的上下文表示 H_{rel} 后，通过一个多层感知机分类模型和 softmax 输出层得到关系的分类概率，计算方式如下：

$$P_{rel} = \text{softmax}(\text{MLP}(H_{rel})) \quad (8)$$

采用基于交叉熵的损失函数作为关系抽取模块的损失函数，计算方式如下：

$$L_{rel} = -\sum_{i=1}^{n}[y^i_{rel}\log P^i_{rel} + (1 - y^i_{rel})\log(1 - P^i_{rel})] \quad (9)$$

2.4 联合学习方法

联合学习过程中，实体识别模块和关系抽取模块共享参数，能够充分利用两个子任务的关联性对预训练模型 BERT 进行调优。整个联合抽取模型的损失函数 L_{joint} 由两个子任务的损失函数（公式（3）和（9））共同决定，最终联合学习的损失函数定义如下，其中 λ 为一个用于平衡实体识别模块损失和关系分类模块的超参数。

$$L_{joint} = \lambda L_{ner} + (1 - \lambda) L_{rel} \quad (10)$$

3 实验与结果分析

3.1 数据集和评价指标

实验部分主要采用 ADE 公开数据集[2]，达到与此前研究可对比的效果。该数据集主要由 5 位独立的领域专家通过共同讨论制定标注指南文件，再由 3 位专家实际进行数据标注得到，具体统计信息，见表 1。评价指标主要由实体识别的评价指标、关系抽取的评价指标和实体关系联合抽取评价指标 3 部分组成。采用机器学习领域常用的精准率、召回率和 *F*1 指数。为了便于与此前研究方法进行性能对比，通过与此前方法类似的 10 折交叉验证来验证模型效果。

表 1 训练集数据统计信息

数据类型	类别	数量（个）
实体	药物	5 063
	副作用	5 776
关系	药物 - 副作用关系	1 551

3.2 实验参数设置

为了比较不同预训练模型在本文设计提出的实体关系联合抽取框架中的实际效果，测试 BERT、BioBERT 和 ClinicalBERT 共 3 种预训练模型的表现。实验中联合抽取模型使用的具体参数，见表 2。

表 2 本文实验中的参数设置

参数	数值
最大序列长度	128
训练批次大小	16
初始化学习率	2×10^{5}
训练时期数	5
丢弃率	0.1
L2 正则化系数	5×10^{-3}

3.3 实验结果

3.3.1 预训练模型对比 实验结果，见表 3、表 4。基于生物医学文献训练得到的 BioBERT 模型在面向生物医学文献中的 ADE 实体和关系抽取时 *F*1 表现（0.904，0.868）明显优于基于书籍语料和维基百科语料训练得到的 BERT，以及基于临床文本训练得到的 ClinicalBERT。但是在端到端任务的验证结果方面，本文方法结合 3 种不同模型时 *F*1 表现则较为接近，见表 5。

表 3 本文方法结合不同预训练模型在实体抽取任务中的实验结果

预训练模型	精准率	召回率	*F*1
BERT	0.868	0.907	0.886
BioBERT	0.891	0.918	0.904
ClinicalBERT	0.849	0.909	0.878

表 4 本文方法结合不同预训练模型在关系抽取任务中的实验结果

预训练模型	精准率	召回率	*F*1
BERT	0.842	0.851	0.847
BioBERT	0.824	0.917	0.868
ClinicalBERT	0.826	0.866	0.845

表 5 本文方法结合不同预训练模型在端到端抽取任务中的实验结果

预训练模型	精准率	召回率	*F*1
BERT	0.807	0.959	0.877
BioBERT	0.801	0.973	0.878
ClinicalBERT	0.799	0.964	0.873

3.3.2 与现有方法对比 本文所设计的基于预训练模型的实体关系联合抽取方法在 ADE 数据集上的实体抽取表现和关系抽取表现（0.904，0.868）均优于此前的研究[6-9,16]，见表 6、表 7。实验数据均来自原作者发表论文。同是基于预训练模型的方法，本文方法在实体识别和关系抽取上的表现均优于 Giorgi J 等[16]提出的方法。端到端任务实验结果，见表 8，本文方法（0.878）与 Giorgi J 等[16]的方法（0.877）表现接近，优于其他现有方法。

表 6 本文方法和现有方法在实体识别任务中的实验结果

方法	精准率	召回率	*F*1
Li F 等[6]	0.795	0.796	0.795
Li F 等[7]	0.827	0.867	0.846
Ramamoorthy S 等[8]	0.884	0.824	0.853
Bekoulis G 等[9]	0.847	0.882	0.864
Giorgi J 等[16]	–	–	0.896
本文方法	0.891	0.918	0.904

表 7 本文方法和现有方法在关系抽取任务中的实验结果

方法	精准率	召回率	*F*1
Li F 等[6]	0.640	0.629	0.634
Li F 等[7]	0.675	0.758	0.714
Ramamoorthy S 等[8]	0.863	0.873	0.868
Bekoulis G 等[9]	0.721	0.772	0.746
Giorgi J 等[16]	–	–	0.858
本文方法	0.824	0.917	0.868

表 8 本文方法和现有方法的端到端实验结果

方法	精准率	召回率	*F*1
Li F 等[6]	–	–	0.715
Li F 等[7]	–	–	0.780
Bekoulis G 等[9]	–	–	0.805
Giorgi J 等[16]	–	–	0.877
本文方法	0.801	0.973	0.878

4 讨论

通过实验分析发现，本文提出的基于预训练模型的实体关系联合抽取方法仍存在一定的改进空间，其中包括实体和关系抽取模块的优化设计、联合学习的方法等。

4.1 模块设计

本文在实体识别模块中采用一种基于预训练模型和前向神经网络的结构，虽然也取得不错的效果，但是对预训练模型的利用仍存在改进空间。后期可以采用已经在某些数据集上验证的更优化的神经网络结构，如 Si Y 等[19]使用 BiLSTM + BERT 的方法进行改进。随着研究者对预训练模型研究的深入，将提出更多的实体抽取或关系抽取方法，本文提出的联合抽取框架具有一定的扩展性，即实体抽取和关系抽取模块能够被更优化的基于预训练模型的方法替换。

4.2 预训练方式

对于预训练模型本身，本文方法并没有处理其预训练过程，而是采用通用方法得到预训练模型。对于预训练过程，可以考虑融合多种新的任务或者方法扩展原有基于掩码的语言模型（masked language model，MLM）和基于下一句预测（next sentence prediction，NSP）的方法，使训练得到的预训练模型在端到端实体关系任务上获得更优表现。现有的预训练模型对于序列分类任务和序列标注任务都设计了有针对性的训练方法，从而得到在多项测试集上的优异结果，但尚无针对关系抽取的特定优化或针对端到端方法对预训练模型本身进行的优化，导致在部分实例上效果不佳。

4.3 联合学习方法

除了利用联合抽取框架平衡两个模块的方法外，Zheng S 等[20]于 2017 年提出标注方法解决实体关系联合抽取问题，即将实体关系联合抽取转换为与实体识别类似的序列标注任务，以“BIEO – Relation – Entity”的形式，将实体信息和关系信息都包含在每个 token 标签中。上述方法虽然存在无法处理实体重叠的问题，但是仍然为研究者打开了一种新的研究思路，多重标注或者多次识别可能弥补上述短板从而衍生出新的实体关系抽取方法。Zheng S 等[20]也在 NYT 数据集上验证了其方法的有效性。

5 结语

本文结合医学自然语言处理领域的最新发展趋势，面向 ADE 抽取任务提出了一种基于预训练模型的实体关系联合抽取方法。充分利用预训练模型在特征表示上的优势，无须人工加入对于实体或者序列的表示特征。实验结果表明，该方法优于已有联合抽取方法，能够应用于 ADE 的抽取中。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 YASREBI - DE KOM I, DONGELMANS D A, DE KEIZER N F, et al. Electronic health record - based prediction models for in - hospital adverse drug event diagnosis or prognosis: a systematic review [J]. Journal of the American medical informatics association, 2023, 30 (5): 978 - 988.

2 吉向敏. 基于数据挖掘与网络模型的药物不良事件预测及监测研究 [D]. 哈尔滨：哈尔滨工程大学，2020.

3 DA SILVA B, KRISHNAMURTHY M. The alarming reality of medication error: a patient case and review of Pennsylvania and National data [J]. Journal of community hospital internal medicine perspectives, 2016, 6 (4): 31758.

4 GURULINGAPPA H, RAJPUT A M, ROBERTS A, et al. Development of a benchmark corpus to support the automatic extraction of drug - related adverse effects from medical case reports [J]. Journal of biomedical informatics, 2012, 45 (5): 885 - 892.

5 HENRY S, BUCHAN K, FILANNINO M, et al. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records [J]. Journal of the American medical informatics association, 2020, 27 (1): 3 - 12.

6 LI F, ZHANG Y, ZHANG M, et al. Joint models for extracting adverse drug events from biomedical text [C]. New York: 25th International Joint Conference on Artificial Intelligence, 2016.

7 LI F, ZHANG M, FU G, et al. A neural joint model for entity and relation extraction from biomedical text [J]. BMC bioinformatics, 2017, 18 (1): 1 - 11.

8 RAMAMOORTHY S, MURUGAN S. An attentive sequence model for adverse drug event extraction from biomedical text [EB/OL]. [2023 - 04 - 10]. https://arxiv.org/pdf/1801.00625.pdf.

9 BEKOULIS G, DELEU J, DEMEESTER T, et al. Adversarial training for multi - context joint entity and relation extraction [C]. Brussels: The 2018 Conference on Empirical Methods in Natural Language Processing, 2018.

10 LI Q, JI H. Incremental joint extraction of entity mentions and relations [C]. Baltimore: The 52nd Annual Meeting of the Association for Computational Linguistics, 2014.

11 MIWA M, SASAKI Y. Modeling joint entity and relation extraction with table representation [C]. Doha: The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

12 MIWA M, BANSAL M. End - to - end relation extraction using lstms on sequences and tree structures [C]. Berlin: The 54th Annual Meeting of the Association for Computational Linguistics, 2016.

13 ADEL H, SCHÜTZE H. Global normalization of convolutional neural networks for joint entity and relation classification [C]. Copenhagen: The 2017 Conference on Empirical Methods in Natural Language Processing, 2017.

14 KATIYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees [C]. Vancouver: The 55th Annual Meeting of the Association for Computational Linguistics, 2017.

15 黄敏婷，赵静，于涛. 基于医学大数据的预训练语言模型及其医学文本分类研究 [J]. 中华医学图书情报杂志，2021，29 (11)：39 - 46.

16 GIORGI J, WANG X, SAHAR N, et al. End - to - end named entity recognition and relation extraction using pre - trained language Models [EB/OL]. [2023 - 04 - 10]. https://arxiv.org/pdf/1912.13415.pdf.

17 MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations ofwords and phrases and their compositionality [C]. Lake Tahoe: The 26th International Conference on Neural Information Processing Systems, 2013.

18 KIM Y, JERNITE Y, SONTAG D, et al. Character - aware neural language models [C]. Phoenix: The Thirtieth AAAI Conference on Artificial Intelligence, 2016.

19 SI Y, WANG J, XU H, et al. Enhancing clinical concept extraction with contextual embeddings [J]. Journal of the American medical informatics association, 2019, 26 (11): 1297 - 1304.

20 ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme [C]. Vancouver: The 55th Annual Meeting of the Association for Computational Linguistics, 2017.