基于深度学习的中医古文献临床经验抽取

卢永美[1]，卜令梅[1]，陈 黎[1]，于中华[1]，张婷婷[2]，叶 莹[3]

(1. 四川大学计算机学院，成都 610065；
2. 成都中医药大学医学信息工程学院，成都 610075；
3. 成都中医药大学基础医学院，成都 610075)

摘 要：中医古文献蕴藏着丰富的临床经验，是古代中医在行医过程中对临床诊疗的经验性总结，体现了中医学形成和发展的理论框架和思想基础. 然而这些宝贵的临床经验不仅量大，而且分散在不同的文献中，使得中医从业者手工很难快速全面地获取它们，文献检索工具也只能提供文档级别的信息筛选，无法为这种细粒度的信息获取提供支持. 此外，古汉语相对于现代汉语的不同特点也限制了主流文本分析工具的使用效果. 为此本文提出面向临床经验获取的中医古文献信息抽取任务，用于识别古文献中描述临床经验的文本片段，手工标注了样本数据用于这种抽取模型的训练和测试，并设计了基于深度学习的序列标注器用于完成该任务. 考虑到标注数据量小可能带来的过度拟合问题，本文引入对抗训练和虚拟对抗训练来增强模型的泛化能力. 一系列充分的实验验证了模型的有效性，表明利用信息抽取技术从古文献获取中医临床经验具有可行性，为这一新的信息抽取任务提供了有希望的研究基线和可复用的标注数据集.

关键词：中医古文献；临床经验；深度学习；序列标注

中图分类号：TP391 **文献标识码**：A DOI：10.19907/j.0490-6756.2022.023005

Extracting clinical experiences from ancient literature of traditional chinese medicine via deep learning

LU Yong-Mei[1], *BU Ling-Mei*[1], *CHEN Li*[1], *YU Zhong-Hua*[1], *ZHANG Ting-Ting*[2], *YE Ying*[3]

(1. College of Computer Science, Sichuan University, Chengdu 610065, China;
2. College of Medical Information Engineering, Chengdu University of TCM, Chengdu 610075, China;
3. College of Basic Medical, Chengdu University of TCM, Chengdu 610075, China)

Abstract: Ancient literature of Traditional Chinese Medicine (TCM) contains rich clinical experiences, which is the empirical summary of clinical diagnosis and treatment in the process of ancient Chinese medicine practice, and embodies the theoretical framework and ideological basis of the formation and development of TCM. However, due to the volume and dispersion of valuable clinical experiences, it is difficult for TCM doctors to quickly and comprehensively obtain the clinical information they need from ancient literature manually, and the document retrieval tools can only provide document-level information screening, which cannot support fine-grained information extraction. In addition, the different characteristics of ancient Chinese relative to modern Chinese also limit the use of mainstream text analy-

收稿日期：2021-09-26
基金项目：国家重点研发项目(2020YFB0704502)；国家自然科学基金(61801058)
作者简介：卢永美(1997—)，女，贵州遵义人，硕士研究生，研究方向为自然语言处理，医学信息学.
通讯作者：于中华. E-mail：yuzhonghua@scu.edu.cn

sis tools. For this reason, we propose a task of information extraction from the ancient literature of TCM for obtaining clinical experiences, which is used to identify text fragments describing clinical experiences in ancient literature and manually annotate sample data for training and testing the extraction task, a sequence labeling model is designed based on deep learning to complete the task. Considering the overfitting problem that can be brought about by the small amount of annotated data, we introduce adversarial training and virtual adversarial training to enhance the generalization ability of the proposed model. A series of sufficient experiments are conducted on the clinical experience dataset to verify the effectiveness of the model, and the experimental results show the feasibility of extracting clinical experiences from ancient literature by information extraction technology, and a promising baseline and a reusable annotated dataset for the new information extraction task are available.

Keywords: Ancient literature of TCM; Clinical experiences; Deep learning; Sequence labeling

1　引　言

中医古文献包含了几千年来中医从业者在临床诊疗中的经验性总结.这些经验总结是中医知识的重要组成部分,对现在的中医临床实践有着重要的指导价值.如图 1 所示,临床经验描述了疾病的症状、用药以及煎服方法等信息,它为现代中医进行各种疾病的临床诊断和治疗提供了大量参考.甚至 2015 年诺贝尔医学奖获得者屠呦呦也是受到了东晋葛洪的《肘后备急方》一书中"青蒿一握,以水二升渍,绞取汁,尽服之"的启发,成功提取出青蒿素并研制出抗疟新药.

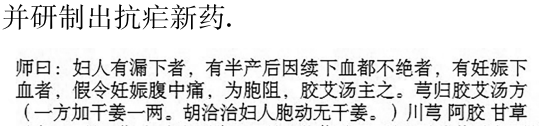
师曰：妇人有漏下者，有半产后因续下血都不绝者，有妊娠下血者，假令妊娠腹中痛，为胞阻，胶艾汤主之。芎归胶艾汤方（一方加干姜一两。胡洽治妇人胞动无干姜。）川芎 阿胶 甘草（各二两） 艾叶 当归（各三两） 芍药（四两） 干地黄（六两）上七味，以水五升，清酒三升，合煮取三升，去滓，纳胶令消尽，温服一升，日三服。不瘥更作。

图 1　临床经验实例

Fig. 1　An example of clinical experience

然而,医学工作者从海量的古文献中手工筛选所需要的临床经验耗时耗力.据不完全统计,目前有 10 000 多种中医古文献,其中有 37 000 多种版本[1].此外,古文献使用的古汉语和现代汉语的语言风格差异很大,如文献中常会出现通假字、古今字等一字多用,"中风"等一词多义和"妊,娠,孕,胎"等多词一义的现象.虽然现有一些检索工具,如《中华医典》,能够辅助医生从古文献中检索临床经验,但是基于字面相似性的全文检索系统依旧存在检索结果噪声大、检索性能不好等问题.因此,这些挑战严重阻碍了研究者从古文献中获取临床治疗经验.古文献对现代中医研究和临床实践的重要性人们早就认识到了,但直到最近,一些研究者才开始利用文本挖掘和信息抽取技术对古文献进行分析处理,如术语规范[2]、方药配伍[3]、医案分类[4,5]和知识图谱构建[6]等.

据我们所知,目前还没有从中医古文献中自动抽取临床经验的相关研究及检索工具,而这样的研究成果对辅助中医临床诊断以及中医的理论研究起着积极的促进作用.因此,本文提出从古文献中自动抽取临床经验文本片段的任务.古文献临床经验的自动抽取能为中医领域下游的研究(如方剂溯源,症状名演变等)提供重要的数据支撑.

本文把从古文献中抽取临床经验(Extraction of Clinical Experiences,ECE)的任务归结为序列标注问题,为了验证 ECE 任务,本文手工构建了数据集,并提出一个序列到序列的深度学习模型.首先,模型使用卷积神经网络(Convolutional Neural Network,CNN)学习句子中字的 n 元组的表示,然后通过最大池化聚合形成句子表示.然后,进一步地利用一个双向长短时记忆网络(Bi-directional Long Short-Term Memory,Bi-SLTM)来聚合句子上下文的特征,从而输出的句子嵌入表达不仅携带了当前句子的信息,也包含了句子间的前后文信息.此外,考虑到标签之间存在的关联性,模型利用条件随机场(Conditional Random Field,CRF)[7]输出优化后的标签序列.

近年来,使用词嵌入的深度学习方法在处理文本方面非常流行.然而,由于文本中词的偏态分布,基于深度学习的方法总是受到未登录词 (Out-Of-Vocabulary,OOV) 的影响.为了克服 OOV 挑战,研究者们提出使用子词(sub-words)嵌入(对于中文来说是字嵌入)来提升文本任务的性能[5,8,9].此外,与处理现代汉语文本不同,处理古汉语文本面临着分词的困境,即由于字词差异模糊而导致的分词困难.因此,对于古汉语文本,使用字嵌入而不是

词嵌入更合理. 因此，在本文模型中，一个句子被认为是由字序列组成，字是最小的处理单元而不是词.

此外，由于临床经验数据集手工标注工作量和时间耗费过于庞大，使得可以获取的标注数据集规模有限. 众所周知，深度学习模型在小数据上容易出现过拟合现象. 为了解决这个问题，本文引入对抗训练(Adversarial Training，AT)[10]和虚拟对抗训练(Virtual Adversarial Training，VAT)[11]两种不同的方法来增强模型的泛化能力，并以此来进一步提高抽取性能.

本文在有限的临床经验数据集上进行了一系列实验，并从两个角度验证临床经验的抽取性能：一是模型在句子级别的分类能力；另一个是模型抽取完整临床经验文本片段的能力. 本文的实验结果表明，在对句子的标注性能可以达到 78.53%的 F_1 值和 81.5%的准确率. 临床经验片段的抽取上，性能可以达到 61.17%的精确率和 51.14%的召回率. 实验结果证明了从古文献中抽取临床经验的任务是可行的，本文提出的模型对抽取任务是有效的.

本文的贡献主要有以下 4 个方面：(1) 提出了古文献的临床经验自动抽取任务，并手工构建了用于训练和测试的临床经验数据集；(2) 将临床经验抽取任务转换为序列标注问题，并提出一个基于深度学习的序列标注模型；(3) 针对数据集规模小的问题，引入对抗训练和虚拟对抗训练两种方法来解决模型的泛化能力问题；(4) 在构建的临床经验数据集上进行了一系列实验，从两个角度验证了抽取任务的可行性和本文模型的有效性.

2　相关工作

信息抽取(Information Extraction，IE)作为自然语言处理的基本任务之一，在医学、材料和法律等各个研究领域得到广泛研究. 在医学领域，IE 常用于提取医学文本(例如电子病历)中的实体以及关系，来帮助构建医学知识图谱，以辅助医生进行医学决策[12-14]. 随着 IE 在医学领域的广泛研究，其也进入了中医的视野. 目前大多数的研究工作都以现代汉语书写的结构化或非结构化文本为研究对象，如方药配伍[15,16]、辨证论治[17]、知识图谱构建[18]和细粒度实体语料库构建[19]等. 古文献作为古中医的文本载体，记载了丰富的中医医学信息，因此，对中医古文献的分析研究有利于发挥中医的原始优势.

近几年，利用机器学习技术对古文献的挖掘和分析逐渐成为研究热点. 2014 年，Weng 等[2]利用隐马尔可夫模型对古文献中与脾有关的短语进行医学术语识别，以此进行脾相关术语规范，并进一步开发了一套系统来支持与脾相关的中医研究. 2015 年，聂佳等[3]利用关联规则算法对巴蜀中医古文献中医案进行数据挖掘与分析，拟在进一步探究巴蜀中医学术流派的辨证施治、用药规律等. 2016 和 2019 年，Yao 等[4,5]提出了中医古文献医案临床分类的任务，分别使用了传统的机器学习技术(例如 SVM、MaxEnt 等)与现下流行的 BERT 语言模型对古文献中的医案进行分类，且获得不错的性能. 2019 年，Zhou 等[6]通过摘录中医学相关资料(包含中医古文献)中与疾病、症状、方剂和药物等相关概念构建了中医药知识图谱，期望形成完善的知识服务体系. 同年，Gao 等[20]综合调查了中医古文献研究进展，提到古文献研究面临着巨大的数据挑战：大部分研究者手动收集、检索和整理数据，这些数据通常会遗漏信息，以及宝贵的知识.

简而言之，利用人工智能技术对中医古文献智能分析研究还处于起步阶段. 据我们所知，古文献中的临床经验的自动抽取研究还属于空白. 为了弥补这一空白，本文提出了一个新的中医古文献信息抽取任务，其中待提取的实体是文献中存在的临床经验，并进一步提出了一种基于深度学习的框架来解决此任务.

3　任务与模型

本节将会描述 ECE 任务并介绍本文提出的模型. 针对任务的特点和挑战，本文将 ECE 任务转换成序列标注问题，并且提出了基于字符的序列到序列模型，同时考虑到标注数据集规模较少，在模型尝试了不同的正则化方法来增强模型的泛化性能.

3.1　任务定义

古文献具有一定的篇章结构，可视为多个小节构成的一个文档. 通常来说，临床经验是由中医古文献小节中的几个连续句子组成，并且几乎没有跨越章节之间的边界. 因此，本文模型的输入以每一小节为单位，并采用了流行的 BIO 策略进行标注.

假设给定一节 $D=(S_1,S_2,\cdots,S_l)$ 由 l 个句子组成，其中每个句子 $S_i=(c_1,c_2,\cdots,c_n)$ 是由 n 个字组成. 本文的任务是为输入的 D 中每个句子 S_i 确定唯一标签 $y_i\in\{'B','I','O'\}$.

3.2　提出的模型

本文提出的 ECE 模型由两层组成：句子编码

层和序列标注层,如图 2 所示. 首先将古文献拆分为多个小节,每个小节 D 作为模型的输入. 句子编码层对 D 中每个句子通过 CNN 来学习句子表达. 然后将所获得的 D 中所有句子的嵌入表达输入到序列标注层,经过 Bi-LSTM 为每个句子获得更为丰富的上下文信息. 最终经过一个前馈神经网络与 CRF 优化句子的标签.

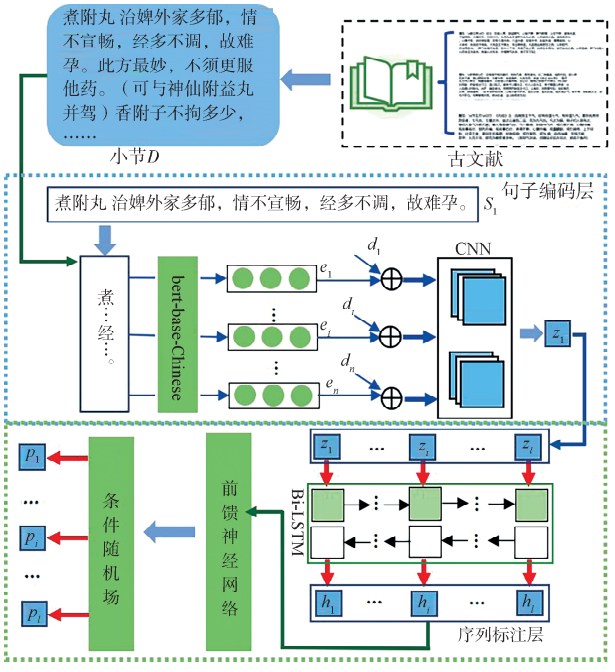

图 2　ECE 模型结构图

Fig. 2　The architecture of ECE model

3.3　句子编码层

每一篇古文献按照篇章结构被切分为多个小节,每一小节 D 由 l 个句子组成. 句子编码层的目标为小节 D 中的每个句子生成一个对应的句子表达. 对于由 n 个字构成的句子 S,本文使用预训练的 bert-base-Chinese 语言模型[21]对 S 中每个字进行初始化嵌入得到对应的字嵌入序列 $e=(e_1,e_2,\cdots,e_n)$. 本文使用 CNN 作为句子编码器,它通过带有卷积滤波器的层来提取字之间的局部依赖关系. 具体来说,对于一个包含 n 个汉字句子 S,可以表示为

$$\hat{e}_{1:n}=\hat{e}_1\oplus\hat{e}_2\oplus\cdots\hat{e}_n \tag{1}$$

其中$\hat{e}_i$是归一化之后的字嵌入(见 3.5 节)和$\oplus$表示连接操作. 令$\hat{e}_{i:i+j}$表示$\hat{e}_i$,$\hat{e}_{i+1}$, …,$\hat{e}_{i+j}$的连接,g 表示一个宽度为 a 的卷积核. 对于一个特征c_i,其由 N-gram($N=a$)窗口$\hat{e}_{i:i+a-1}$产生:

$$c_i^g=\tanh(g\cdot\hat{e}_{i:i+a-1}+b) \tag{2}$$

其中,b 是偏置项. 注意本文使用了填充操作. 所以对于卷积核 g 和长度为 n 的句子 S,可通过式(3)获得序列长度为 n 的 N-gram 向量表示.

$$c^g=(c_1^g,c_2^g,\cdots,c_n^g) \tag{3}$$

最后,对于序列c^g,本文使用最大池化得到句子编码 z.

3.4　序列标注层

D 经过句子编码层可以获得一个向量序列$(z_1,z_2,\cdots,z_l)$,其与句子序列一一对应. Bi-LSTM 考虑了句子前后的上下文信息,可以捕获长距离依赖关系和双向语义信息. 因此,为了进一步考虑小节 D 中句子之间的关系,本文使用 Bi-LSTM 作为编码器. 对于第 i 个句子的句子表达z_i,Bi-LSTM 通过拼接每个时刻前向和后向的隐藏表达获取句子表达h_i,如式(4)所示.

$$h_i=[\overrightarrow{LSTM}(h_i,z_i);\overleftarrow{LSTM}(h_i,z_i)] \tag{4}$$

获得最终的句子表达$(h_1,h_2,\cdots,h_l)$之后,将其作为输入送入到前馈神经网络中来得到每个句子属于每个标签的概率分布,本文用r_i来表示句子S_i相应的概率向量.

最后,本文使用 CRF 优化标签序列. 它可以对标签之间的依赖性进行建模,从而进一步优化标签序列. 对于 D 包含的 l 个句子$S_1,S_2,\cdots,S_l$和其对应的句子标签概率分布为$r_1,r_2,\cdots,r_l$,CRF 的目标是通过计算所有可能标签序列的得分为其获得得分最高的最优标签序列$y^*=(y_1^*,y_2^*,\cdots,y_l^*)$. 对于一个可能标签序列 $y=(y_1,y_2,\cdots,y_L)$,所获得分通过式(5)计算,其中r_{i,y_i}表示句子S_i标注为y_i的概率,和$A_{y_iy_{i+1}}$表示标签y_i后面跟随y_{i+1}标签的转移概率(其在训练阶段学习).

$$\text{score}(y)=\sum_{i=1}^{l}r_{i,y_i}+\sum_{i=1}^{l-1}A_{y_iy_{i+1}} \tag{5}$$

然后,利用 softmax 将 D 的所有可能标签序列 Y 的得分转换为这些标签序列的概率分布,如式(6)所示.

$$P(y)=\frac{e^{\text{score}(y)}}{\sum_{\hat{y}\in Y}e^{\text{score}(\hat{y})}} \tag{6}$$

然后,使用维特比算法为 D 选择最优标签序列y^*,如式(7)所示.

$$y^*=\underset{\hat{y}\in Y}{\text{argmax}}\ \text{score}(\hat{y}) \tag{7}$$

本文通过使用最小化负 log 似然与 L_2 正则损失进行联合训练,如公式(8)所示. 其中 ω 表示所有模型参数,λ 是超参数.

$$L_{\text{origin}}=\log\Big(\sum_{\hat{y}\in Y}e^{\text{score}(\hat{y})}\Big)-\text{score}(y)+\lambda\|\omega\|_2 \tag{8}$$

3.5　对抗训练和虚拟对抗训练

深度学习模型容易在小数据上过拟合，正则化在深度学习中防止过拟合十分有效. 本文将对抗训练和虚拟对抗训练作为一种有效的方法，训练时在字嵌入上添加小扰动来正则化分类器，以此来增强本文提出的模型的泛化能力[10,11]. 然而，ECE 模型可以学习大范数的嵌入，这使得小范数的对抗性扰动的影响变得微不足道[10]. 为了避免这种影响，本文对字嵌入进行归一化，如式(9)～式(11)所示.

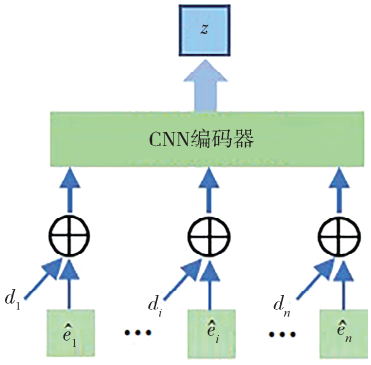

图 3　引入扰动的 ECE 句子编码器

Fig. 3　The sentence encoder of ECE model with perturbation

$$\hat{e}_i=\frac{e_i-E(e)}{\sqrt{D(e)}} \tag{9}$$

$$E(e)=\frac{1}{n}\sum_{i=1}^{n}e_i \tag{10}$$

$$D(e)=\frac{1}{n}\sum_{i=1}^{n}(e_i-E(e))^2 \tag{11}$$

在对抗训练的过程中(如图 3)，在句子编码层归一化之后的字嵌入$\hat{e}=(\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_n)$上加入对抗扰动$d^{AT}$. 在给定当前模型参数$\theta$的情况下，对于对抗扰动$d^{AT}$可通过式(12)～式(13)计算，其中$\varepsilon^{AT}$控制扰动的$L_2$-范数的范围. 对抗损失如式(14)所示计算，其中 K 是标签数. 因为需要训练样本的真实标签分布，所以对抗训练只适用于有监督学习.

$$d^{AT}=\varepsilon^{AT}\frac{g}{\|g\|_2} \tag{12}$$

$$g=\nabla_{\bar{e}}\log p(y|\bar{e};\theta) \tag{13}$$

$$L_{AT}=-\frac{1}{K}\sum_{k=1}^{K}\log p(y_k|\bar{e}_k+d_k^{AT};\theta) \tag{14}$$

本文也使用了虚拟对抗来解决数据集规模小的问题. 与对抗训练不同，虚拟对抗训练并不需训练样本的标签真实分布，所以即使训练样本是没有真实标记的样本点，同样可以加入训练，因此 VAT 不但适用于有监督学习，还适用于半监督学习. 对于虚拟对抗训练，本文使用式(15)和式(16)计算虚拟对抗扰动d^{VAT}，其中 o 是一个很小的随机向量和$KL[p||q]$表示概率分布 p 和 q 之间的 KL 散度. 虚拟对抗训练的损失如式(17)所示计算，其中K'是句子数量.

$$d^{VAT}=\varepsilon^{VAT}\frac{g}{\|g\|_2} \tag{15}$$

$$g=\nabla_{\bar{e}+o}KL[p(\varepsilon|\bar{e};\theta)||p(\varepsilon|\bar{e}+o;\theta)] \tag{16}$$

$$L_{VAT}=-\frac{1}{K'}\sum_{k=1}^{K'}KL[p(\varepsilon|\bar{e};\theta)\|p(\varepsilon|\bar{e}+d_k^{VAT};\theta)] \tag{17}$$

在训练时，ECE 模型的总损失函数定义为

$$L=L_{\text{origin}}+L_A \tag{18}$$

训练使用对抗损失时$L_A=\xi_1 L_{AT}$；使用虚拟对抗训练时$L_A=\xi_2 L_{VAT}$；同时使用时$L_A=\xi_1 L_{AT}+\xi_2 L_{VAT}$. 其中$\xi_1$和$\xi_2$分别是控制对抗训练损失和虚拟对抗训练损失的超参数.

4　实验与结果

4.1　实验数据

本文收集了 368 本中医古籍(不包含医案类)，对所有古籍按照章节结构拆分，得到 4 万多小节. 考虑到需要小节数据量庞大且全部人工标注费时费力，所以本文随机抽取了 1000 节并邀请两名中医专家进行临床经验标注. 一个临床经验主要包含疾病的症状和疗法(方剂或药物组成)，如图 1 所示. 对于标注不统一的情况，两名专家再次进行协商讨论，确定最后的标注结果. 最终手工标注数据集统计结果如表 1 所示.

表 1　实验数据统计

Tab. 1　Statistics of the datasets

小节	临床经验	句子	标签		
			B	*I*	*O*
1000	6641	59 832	6641	23 562	29 629

本文按照 8∶1∶1 比例随机划分了训练集、验证集和测试集，采用了十折交叉验证，并计算了 95%的置信区间. 我们从两个不同的角度评价临床经验抽取的有效性：一是句子的分类性能；另一个是完整临床经验片段的抽取性能. 对于分类性能，本文使用的评价指标包括精准率、召回率、F_1值和准确率. 对于抽取性能，本文使用预测的临床经验的精准率，如式(19)、黄金标准的临床经验的召回率，如式(20)和临床经验的 F_1值，如式(21). 值得注意的是只有当预测的临床经验片段与黄金标准

完全一致才认为是正确的临床经验.

$$P_{ECE}=\frac{\text{预测正确的临床经验}}{ECE\text{ 模型预测的临床经验}} \tag{19}$$

$$R_{ECE}=\frac{\text{预测正确的临床经验}}{\text{黄金标准的临床经验}} \tag{20}$$

$$F_{1_{ECE}}=\frac{2\times P_{ECE}\times R_{ECE}}{P_{ECE}+R_{ECE}} \tag{21}$$

4.2 实　现

本文实现并对比分析了 4 个模型:(1) Baseline:句子编码层编码器为 CNN,序列标注层编码器为 Bi-LSTM 和 CRF 的模型;(2) Baseline-AT:在 Baseline 模型中加入对抗训练,以增强模型的鲁棒性.这是一种有监督学习的方法;(3) Baseline-VAT:在 Baseline 模型中引入虚拟对抗训练.由于计算损失时不需要训练样本的真实标签信息,这是一种半监督学习的方法;(4) Baseline-AT-VAT:在 Baseline 模型中同时引入 AT 和 VAT.这是一种半监督学习方法.

本文使用 bert-base-Chinese(https://github.com/google-research/bert)初始化字嵌入,维度大小为 768.CNN 使用的卷积核大小分别为 3、4 和 5,维度都是 300 维.Bi-LSTM 正向和反向的维度都是 150 维.在训练阶段,本文利用 Adam 优化器[22]学习模型参数,为了避免模型过拟合 dropout 设置为 0.5.本文的学习率设置为 0.0001,L_2 正则项的 λ 设置为 0.0001,对抗训练中 ε^{AT} 和 ε^{VAT} 分别设置为 8 和 4,ξ_1 和 ξ_2 分别设置为 0.2 和 0.05.对于有监督学习和半监督学习,训练集、验证集和测试集使用的同等数据样本;其中使用 VAT 部分时不使用训练样本的真实标签信息.

4.3 句子分类结果分析

本文在临床经验数据集上的句子分类结果如表 2 所示,Baseline 模型在 F_1 值和 Acc 上分别达到 77%和 79.7%,表现出不错的分类效果.同时,引入对抗训练或虚拟对抗训练都显著提高了模型的性能,在对抗训练上提高了 1.53%的 F_1 值和 1.8%的准确率.这样的结果说明加入扰动能有效解决模型的泛化性能并进一步提升模型的性能.另一方面,同时加入对抗训练和虚拟对抗训练的 Baseline-AT-VAT 模型性能与 Baseline-AT 相比,F_1 值降低 0.16%而准确率提升 0.3%;与 Baseline-VAT 相比,F_1 和准确率明显提升 1.24%和 1.4%.这进一步表现出引入对抗训练比引入虚拟对抗训练更有优势.

总体来看,引入对抗训练获得更佳的实验结果.虽然单独引入 VAT 带来的改进并不优于引入 AT,但 VAT 的优点是可以使用无标签数据,这使得利用与有标签数据来自同一来源的大量未标记数据能更好地泛化模型的可能性.

4.4 临床经验抽取结果分析

为了分析 ECE 模型在抽取完整临床经验的效果,本文在临床经验数据集上计算并统计了实验结果,如表 3 所示.

表 2　句子分类实验结果

Tab.2　The results of sentence classification

Model	P/%	R/%	F_1/%	Acc/%
Baseline	78.87±2.66	75.97±2.22	77±2.51	79.7±2.89
Baseline-AT	**80.63±2.85**	**77.43±2.63**	**78.53±2.73**	**81.2±3.05**
Baseline-VAT	79.23±2.57	75.97±2.19	77.13±2.34	80.1±2.79
Baseline-AT-VAT	80.5±2.47	77.07±2.49	78.37±2.48	81.5±2.82

表 3　临床经验文本片段抽取实验结果

Tab.3　The results of extraction of clinical experiences

Model	P_{ECE}/%	R_{ECE}/%	F_{1ECE}/%
Baseline	57.65±3.41	47.92±3.88	52.22±3.48
Baseline-AT	**61.17±5.35**	**51.14±3.77**	**55.48±4.05**
Baseline-VAT	59.26±4.77	48.38±4.27	53.06±3.99
Baseline-AT-VAT	60.14±5.00	49.34±4.26	54.07±4.38

从表 3 实验结果可以看出,Baseline-AT 同样获得了最佳实验结果,61.17%的 P_{ECE}、51.14%的 R_{ECE} 和 55.48%的 $F_{1_{ECE}}$.将表 2 与表 3 进行比较,可以看到识别完整临床经验的性能大幅下降,这体现了本文提出新的信息提取任务中的特殊困难.与其他信息提取任务中要提取的实体不同,本文要提

取的临床经验在古文献中往往具有更大的跨度，而且在古文献中，临床经验总是稀疏出现. 所有这些现象使本文的框架不得不面对机器学习中未解决的三个挑战，即数据分布偏斜、数据稀疏性和序列标记中的长距离依赖.

此外，表 3 结果体现出本文提出的模型框架至少能召回一半的完整临床经验，以及在预测的临床经验中 60%以上的都是正确的. 这证明本文模型对临床经验的抽取是有效的并且可行的. 而且，还可以观察发现相较于 Baseline，引入 AT 和 VAT 的模型性能都得到了显著的提升，更进一步说明了引入 AT 和 VAT 能有效提升模型性能.

4.5　错误分析

此外，本文独立地对每种类型标签的分类性能进行了进一步的研究. 本文选择了十折交叉验证结果其中一折的结果作为分析目标，其中 $B:I:O$ 句子比例为 934 : 3317 : 3508. 如图 4 所示，本文对 4 个对比模型的标签分类比率进行了可视化，用不同的阴影表示本文的模型预测的标签类别百分比，其中横坐标表示黄金标准标签，纵坐标代表模型的预测标签. 每个矩阵中的斜对角线分别对应 B、I 和 O 的精度.

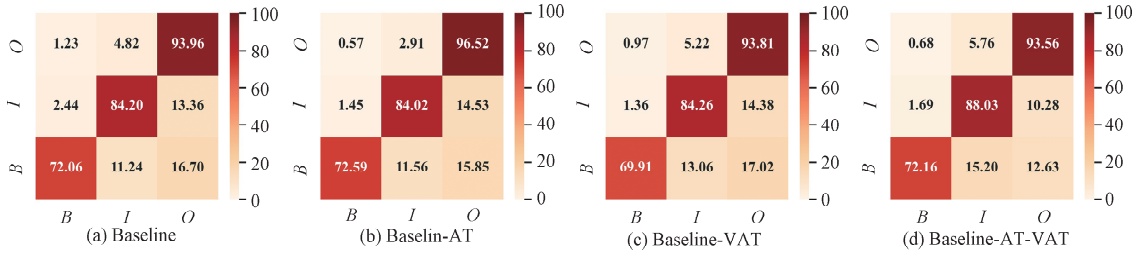

图 4　可视化展示

Fig. 4　Visualized demonstration

图 4 清晰地反应了 BIO 三类标签的分类情况，其能用于比较预测标签和正确标签. 结果表明，BIO 三类标签都产生了较高的精度. 其中 O 标签的精度明显高出 B 和 I 标签，然而 O 与 I 的数量相近，产生这样的结果极有可能是受到 B 标签的影响. B 标签作为临床经验的起始句，这在抽取临床经验任务中起着决定性作用，这也对 I 标签的精度会产生了极大的影响，因为 I 标签始终是出现在 B 标签之后. B 标签的分类性能最高达到 72.59%，能够较好地抽取临床经验的开始边界，这进一步表明模型的可行性与有效性. 对于临床经验边界的识别，本文也在后续研究中对这一事件十分关注.

5　结　论

本文提出了一个新的信息抽取任务，即抽取的实体是中医古文献中的临床经验，以帮助从事中医人员在大量古文献中获取有价值的疾病临床经验. 为此，本文考虑了文献的篇章结构与临床经验的文本片段特点，将临床经验的抽取任务转换为文本片段的序列标注任务. 本文提出了一个基于深度学习的序列标注模型解决该任务. 本文使用 bert-base-Chinese 初始化字嵌入，利用 CNN 作为句子编码器获取 N-gram 信息丰富句子语义编码，并引入文档级别的 Bi-LSTM 学习句子之间的上下文信息进一步丰富句子的语义编码；最后考虑到句子标签之间的关联性，加入 CRF 进行序列标签优化，为每个句子选择最优标签. 为验证提出模型的有效性，本文在专家标注的临床经验数据集上进行一系列实验. 实验验证了本文模型的有效性与可行性. 但也证明了任务特有的困难，特别是在确定临床经验的确切跨度和解决发生稀疏性问题方面，这是本文今后对这一新信息提取任务的学习方向.

参考文献：

[1]　刘耀，周扬. 中医药古文献语料库词语标识标准探讨[J]. 中国中医药信息杂志，2002，9：85.

[2]　Weng H，He W，Ou A，*et al*. Ancient medical literature semantic annotation using hidden markov models [C]//Proceedings of the 2014 IEEE International Conference on Bioinformatics and Biomedicine. Belfast：IEEE，2014.

[3]　聂佳，任玉兰，江蓉星，等. 巴蜀中医药古籍医案数据挖掘系统构建及应用[J]. 中国中医药图书情报杂志，2015，4：13.

[4]　Yao L，Zhang Y，Wei B，*et al*. Traditional Chinese medicine clinical records classification using knowledge-powered document embedding[C]// Proceedings of the 2016 IEEE International Conference on

Bioinformatics and Biomedicine. Shenzhen: IEEE, 2016.

[5] Yao L, Jin Z, Mao C, *et al*. Traditional Chinese medicine clinical records classification with BERT and domain specific corpora [J]. J Am Med Dir Assoc, 2019, 26: 1632.

[6] Zhou Y, Qi X, Huang Y, *et al*. Research on construction and application of tcm knowledge graph based on ancient chinese texts [C]// Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Thessaloniki: ACM, 2019.

[7] Lafferty J, Mccallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proceedings of the Eighteenth International Conference on Machine Learning. Melbourne: Morgan Kaufmann, 2001.

[8] Lee J, Cho K, Hofmann T. Fully character-level neural machine translation without explicit segmentation [J]. T Assoc Comput Linguist, 2017, 5: 365.

[9] Li X, Meng Y, Sun X, *et al*. Is Word segmentation necessary for deep learning of chinese representations? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019.

[10] Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semi-supervised text classification[C]//Proceedings of the 5th International Conference on Learning Representations. Toulon: OpenReview. net, 2017.

[11] Miyato T, Maeda S I, Koyama M, *et al*. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE T Pattern Anal, 2019, 41: 1979.

[12] Liu L, Wu X, Liu H, *et al*. A semi-supervised approach for extracting TCM clinical terms based on feature words [J]. BMC Med Inform Decis, 2020, 20: 118.

[13] Hussain M, Choi D J, Lee S. Semantic based clinical notes mining for factual information extraction [C]//Proceedings of the 2020 International Conference on Information Networking. Barcelona, Spain: IEEE, 2020.

[14] 曹依依, 周应华, 申发海, 等. 基于 CNN-CRF 的中文电子病历命名实体识别研究[J]. 重庆邮电大学学报:自然科学版, 2019, 31: 869.

[15] 张思原, 刘兴隆, 姚攀, 等. 利用稀疏表达学习挖掘中医方剂功效配伍[J]. 四川大学学报: 自然科学版, 2018, 55: 1180.

[16] Chen L, Liu X, Zhang S, *et al*. Efficacy-specific herbal group detection from traditional chinese medicine prescriptions via hierarchical attentive neural network model[J]. BMC Med Inform Decis, 2021, 21: 66.

[17] Xu Q, Tang W, Teng F, *et al*. Intelligent syndrome differentiation of traditional chinese medicine by ANN: a case study of chronic obstructive pulmonary disease[J]. IEEE Access, 2019, 7: 76167.

[18] Li L, Wang P, Yan J, *et al*. Real-world data medical knowledge graph: construction and applications [J]. Artif Intell Med, 2020, 103: 101817.

[19] Zhang T, Wang Y, Wang X, *et al*. Constructing fine-grained entity recognition corpora based on clinical records of traditional Chinese medicine [J]. BMC Med Inform Decis, 2020, 20: 64.

[20] Gao L, Jia C H, Wang W. Recent advances in the study of ancient books on traditional chinese medicine [J]. World J Tradit Chin Med, 2020, 6: 61.

[21] Devlin J, Chang M W, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. Minneapolis: ACL, 2019.

[22] Kingma D, Ba J. Adam: A method for stochastic optimization[EB/OL]. (2017-01-30)[2021-07-22]. https://arxiv.org/pdf/1412.6980v9.pdf.

引用本文格式:

中　文: 卢永美, 卜令梅, 陈黎, 等. 基于深度学习的中医古文献临床经验抽取[J]. 四川大学学报: 自然科学版, 2022, 59: 023005.

英　文: Lu Y M, Bu L M, Chen L, *et al*. Extracting clinical experiences from ancient literature of traditional chinese medicine via deep learning [J]. J Sichuan Univ: Nat Sci Ed, 2022, 59: 023005.