计算机科学
COMPUTER SCIENCE

主实体增强型层叠指针网络在中文医学实体关系抽取中的应用

姜植瀚，咎红英，张莉

引用本文

姜植瀚，咎红英，张莉．主实体增强型层叠指针网络在中文医学实体关系抽取中的应用[J]．计算机科学，2024，51(6A)：230800179-6.

JIANG Zhihan, ZAN Hongying, ZHANG Li. Application of Subject Enhanced Cascade Binary Pointer Tagging Framework in Chinese Medical Entity and Relation Extraction [J]. Computer Science, 2024, 51(6A): 230800179-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800179

主实体增强型层叠指针网络在中文医学实体关系抽取中的应用

姜植瀚[1]　昝红英[2]　张　莉[3]
1 吉林大学软件学院　长春 130012
2 郑州大学计算机与人工智能学院　郑州 450001
3 吉林大学生命科学学院　长春 130012
(2625190314@qq.com)

摘　要　随着中国医学事业的快速发展,中文医学文本的数量不断增加。为了从这些中文医学文本中提取有价值的信息,并解决中文医学领域的实体关系抽取问题,研究人员已经提出一系列基于双向 LSTM 的模型。然而,由于双向 LSTM 的训练速度等问题,文中引入了层叠指针网络框架来处理中文医学文本的实体关系抽取任务。为了弥补层叠指针网络框架中主实体识别能力不足以及解决复用编码层时的梯度问题,文中提出了主实体增强模块,并引入了条件层归一化方法,从而提出了面向中文医学文本的主语增强型层叠指针网络框架(Subject Enhanced Cascade Binary Pointer Tagging Framework for Chinese Medical Text,SE-CAS)。通过引入主实体增强模块,能够精确识别有效的主实体,并排除错误实体。此外,还使用条件层归一化方法来替代原模型中的简单相加方法,并将其应用于编码层和主实体编码层。实验结果证明,所提模型在 CMeIE 数据集上取得了 5.73%的 F1 值提升。通过消融实验证实,各个模块均能带来性能提升,并且这些提升具有叠加效应。

关键词:实体关系抽取;层叠指针网络;医学关系抽取;深度学习;主语识别
中图分类号　TP391

Application of Subject Enhanced Cascade Binary Pointer Tagging Framework in Chinese Medical Entity and Relation Extraction

JIANG Zhihan[1],ZAN Hongying[2] and ZHANG Li[3]
1 Collage of Software,Jilin University,Changchun 130012,China
2 School of Computer and Artificial Intelligence,Zhengzhou University,Zhengzhou 450001,China
3 Collage of Life Science,Jilin University,Changchun 130012,China

Abstract　With the rapid advancement of China's biomedical industry,the volume of Chinese medical texts is escalating at a rapid pace. Extracting valuable information from these texts can ease the learning curve for practitioners. To tackle the challenge of entity relation extraction in the realm of Chinese medicine,a series of models based on bidirectional LSTM have been previously proposed. However,to overcome the training speed bottleneck inherent to bidirectional LSTM,this study introduces the Cascade binary pointer network framework to the domain of Chinese medical filed. To address the framework's weak capability in identifying main entities and the gradient issues arising from reusing the coding layer,this paper introduces the main entity enhancement module and employs conditional layer normalization. This paper presents the subject enhanced cascade binary pointer tagging framework for chinese medical text (SE-CAS),tailored for Chinese medical text. The subject enhancement module accurately identifies valid subjects detected by the subject recognition module and rectifies erroneously identified entities. Furthermore,the conditional layer normalization method replaces the simplistic addition between word embeddings and subject embeddings found in the original model. Experimental results demonstrate that the proposed model achieves a 5.73% enhancement in F1 measure on the CMeIE dataset. The ablation study confirms the incremental impact of each module,and these improvements exhibit a cumulative effect.

Keywords　Entity relation extraction,CASREL,Medical relation extraction,Deeplearning,Subject recognition

1　引言

实体关系抽取任务是自然语言处理中的重要任务之一,指从非结构化或半结构化的文本信息中提取由主实体和客实体构成的实体对以及实体之间的相关关系,并以三元组的形式表示。在生物医学领域,存在着大量非结构化或半结构化的文本信息,相关从业者疲于阅读与日俱增的医学文本,因此面向生物医学领域的实体关系抽取具有重大的研究价值和意义。

根据医学文本中包含的关系数量,可以将任务划分为单关系抽取和多关系抽取。单关系抽取顾名思义,即为需要抽取的句子中只包含一个关系三元组。而多关系抽取则是在一句抽取句子中存在着多重关系,关系抽取中的关系大多为偏序关系,由偏序关系的自反性和结合律可知多重关系都可以被表示为多个关系三元组。这些三元组之间可能会存在多种不同的重叠类型,如同一对实体间有多种不同的关系,或实体之间存在重叠。三元组实体重叠问题又可分为普通型(Normal)、实体对重叠型(Entity Pair Overlap,EPO)和单实体重叠型(Single Entity Overlap,SEO)。普通型即为所有实体对

通信作者:昝红英(iehyzan@zzu.edu.cn)

不存在重叠问题，实体对重叠表示相同实体对间存在着不同关系，单实体重叠型表示两个及以上实体对之间存在着单个实体的重叠关系。

Wei 等[1]提出了基于参数共享的联合实体关系抽取方法，后被称为层叠指针网络框架。其核心在于首先识别可能的主实体，再基于主实体同时识别关系和被关系支配的实体——客实体。但在主实体识别任务中，Wei 将所有识别出的实体都作为主实体去预测客实体。然而，在生物医学领域，主实体的类型是有限的，主要集中在“疾病”这一类别中[2]。在抽取任务中，如果能识别主客实体，则能够快速构建生物医学信息。然而，层叠指针网络框架却无法很好地识别目标主实体，实质上识别出的是具有主实体特征的全部实体。这其中包含着大量的“症状”“治疗”等无关实体。

如图 1 所示，对于给定的输入文本“另外，本病可能发生致死性并发症，即巨噬细胞活化综合征（MAS），其临床表现主要以发热、肝脾淋巴结增大、全血细胞减少、肝功能急剧恶化、凝血功能异常以及中枢神经系统表现为特征，重者甚至发生急性肺损伤及多脏器功能衰竭。骨髓穿刺活检可见吞噬血细胞现象。”，应当识别出的主实体为：“巨噬细胞活化综合征（MAS）”，但在 Wei 等的实验中通常会把“发热”“肝脾淋巴结增大”“全身血细胞减少”“肝功能急剧恶化”和“骨髓穿刺活检”等一系列“症状”和“治疗方法”识别为主实体，降低了后续对关系及客实体解码操作的准确性。

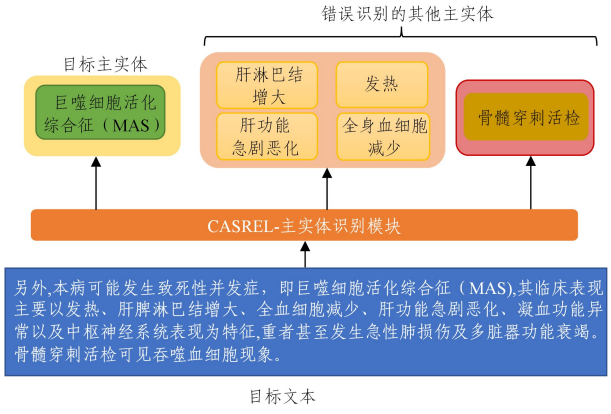

图 1 传统主实体识别示意图

Fig. 1 Diagram of traditional subject recognition

因此，本文提出了面向中文医学文本的主语增强型层叠指针网络模型（Subject Enhanced Cascade Binary Pointer Tagging Framework for Chinese Medical Text，SE-CAS），该模型增强了主语识别的能力，改进了原模型在主语识别过程中出现的冗余以及错误，通过条件归一化层增强了客实体关系的识别能力，SE-CAS 在生物医学领域有着更好的应用效果。SE-CAS 在中文医学信息抽取数据集上取得了良好的实验效果，在精确率、召回率和 F1 值上均取得了突破。其中精确率高达 68.27%，相比传统模型提高 5.81%；F1 值达到 65.55%，相比原方法提高 5.73%。

2 相关工作

2.1 实体关系抽取的发展

早期对关系三元组提取的研究基于流水线（Pipeline）方法，该方法将任务分为两个子任务：命名实体识别和关系分类。首先将问题当作一个简单的命名实体识别任务进行实体的识别，然后根据识别出的实体去研究它们之间的关系。基于 RNN，CNN 和 LSTM 的流水线关系抽取模型被陆续提出。但由于两个子任务之间缺少交互，忽略了关系三元组成员之间的内部关联，以及上游错误的累积传播影响到下游关系分类任务的执行。同时，由于流水线方法直接抽取实体，缺少对同样实体不同关系的考量，无法很好地解决实体重叠问题。

针对这一问题，后续研究利用基于特征工程的联合抽取方法和基于神经网络的联合抽取方法，取得了进展。然而，基于特征工程的方法严重依赖特征的手工构建，需要大量的体力劳动。早期的联合提取方法只是共享神经网络中的权重，但它们仍然独立解码实体和关系。Zheng 等[3]将关系抽取任务转换为序列标注任务，实现了三元组的联合提取。Lee 等[4]于 2018 年提出了一种多层次注意力的远程监督抽取模型，召回率和精确率相比主流方法均有提高。此后，联合实体关系抽取方法发展迅速，涌现出了大量的联合抽取模型，如使用复制策略（Copy Strategy）的端到端模型[5]、融合图卷积神经网络（Graph Convolution Neural Network，GCN）的端到端模型[6]、引入强化学习策略的 Seq2Seq 模型[7]以及具有复制机制（Copy Mechanism）的多任务学习端到端模型[8]，这些模型在一定程度上都取得了较为优秀的结果。后续，随着预训练模型的兴起，越来越多的联合抽取方法引入了 BERT[9]等预训练模型。Wei 等将关系视为从头实体到尾实体的映射函数，并通过名为 CASREL 的二层标记框架完成了三元组联合提取的任务。此外，Wang 等[10]引入握手标记方案来解决曝光偏差的新模型 TPLinker 也成为了实体关系抽取任务中的新范式。Zhang 等[11]为改进 CASREL 的识别效率，提出了一种优先抽取关系然后抽取实体的模型，在 F1 值上取得了进步，但是在客实体识别时，仍然采用了主实体和编码表示向量简单相加的方法。Zhu 等[12]为解决联合学习方法中的嵌套实体和曝光偏差，于 2022 年改进 CASREL 提出了 ATM-REL，但该方法仍然没能解决 CASREL 在主语识别过程中的其他实体识别问题，在主实体识别时仍然会产生冗余。这些方法都没有实现对该模型重用主实体信息可能导致的梯度问题做出改进，也没有对主实体识别造成的巨大浪费做出优化，这两个问题成为限制该模型后续发展的主要屏障。

2.2 生物医学领域的关系实体抽取任务

早在 2011 年 Uzuner 等[13]就将疾病实体的关系抽取添加到电子病历中，考虑了以下 3 种关系：疾病与疾病、疾病与检查、疾病与治疗药物等。在此后的测评中，又陆续加入了以下几种任务：疾病实体与时间之间的关系抽取，以及电子病历中可能导致心脏疾病的风险因素的抽取。2015 年 Wei 等[14]又提出了从生物医学文献中，抽取化学药品和疾病之间可能存在的相关关系的任务，该任务被分为疾病的命名实体识别阶段和化学药物诱导疾病关系提取（Chemical Induced Diseases Relation Extraction，CID）阶段。此后的研究关注于医学文本的关系抽取任务，Sahu 等[15]于 2016 年将 CNN 引入到了医学领域，证实了深度学习方法在医学领域内的可行性。2018 年 Zhang 等[16]将结合 RNN 和 CNN 的深度学习方法引入到生物医学领域的相关信息抽取任务中。Liu 等[17]将 BiLSTM-CRF 模型应用到了医学名词的命名实体识别任务中，在多个数据集上取得了较高的准确率。随着预训练模型的兴起，Luo 等[18]提出了基于注意力的 ATT-BiLSTM-CRF 模型。2020 年，Lee 等[19]提出了一个专为医学文本挖掘的预训练模型 BioBERT，其被作为专用的生物医学模型，在多个数据集上取得了优异的结果。

Zhang 等[20]于 2021 年提出了 BLSTM-MCatt-CNN，尽管在医学领域取得了良好的效果，但是编码器模型仍然使用双向 LSTM，效率不高。目前来看，大量的医学领域的编码器模型仍然集中在双向 LSTM 架构上，LSTM 的训练成本问题以及梯度问题成为限制医学领域实体关系抽取的主要问题。

综上，为了解决医学领域编码向量在预训练模型上训练效率过低的问题，对 CASREL 的主语识别和主实体复用问题做出优化和改进，本文提出了面向中文医学文本的主语增强型层叠指针网络模型，通过主实体增强模块，加强对主实体的识别能力，并且通过引入条件归一化层更好地复用主实体编码，同时解决了上述问题。

3 主语增强型层叠指针网络模型

层叠指针网络的提出，主要是为了解决目前现有模型处理重叠关系三元组效果不好等问题。该模型采用了一种层次化的二级标注框架。首先，从输入句子中检测出主实体；然后，遍历选出的所有主实体，对每个主实体在句子中寻找对应关系和客实体，尝试在句子中提取寻找到的三元组。与这两个步骤相对应，解码器由主实体解码模块（Subject Tagger）和客实体-关系解码模块（Relation-specific Object Taggers）两部分组成。主实体解码模块通过直接解码编码器产生的编码表示向量来识别输入句子中所有可能的主实体，客实体-关系解码模块通过输入编码层和主实体信息的组合同时识别尾实体以及该客实体与相应主实体构成的关系。

针对传统的层叠指针网络模型不能识别特定主实体的问题，本文模型提出了主实体增强模块。同时，采用了条件层归一化的方式取代了原模型中句子编码信息与主实体简单相加的方式，提高了模型的运行效率和准确率。为了使模型在特定场合有着更好的效果，本文根据中文医学信息抽取数据集做了相应的改进。该模型的架构如图 2 所示。

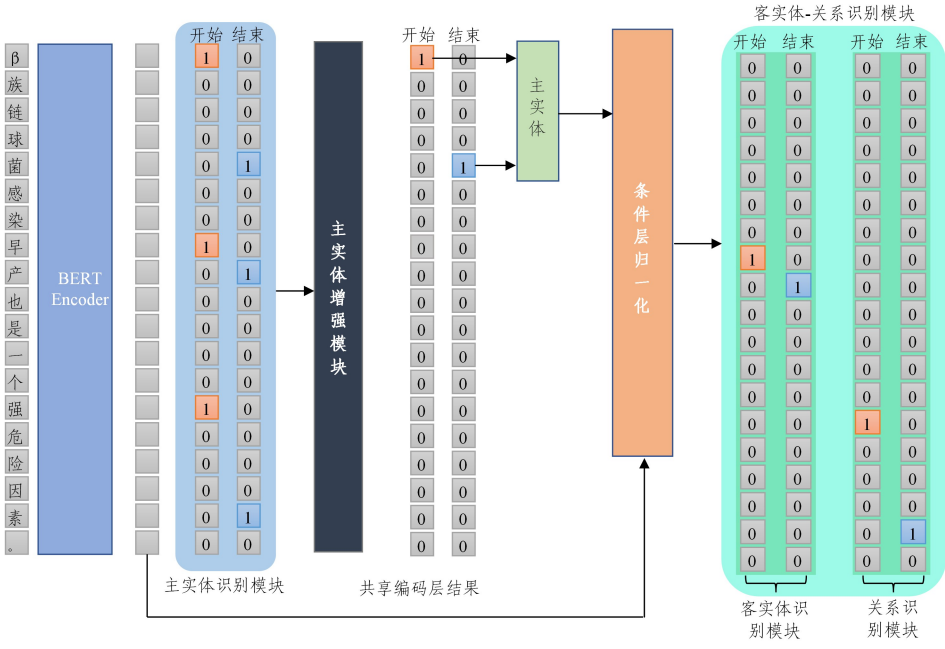

图 2 SE-CAS 模型的架构

Fig. 2 Schema of SE-CAS

3.1 编码模块

为解决在医学领域传统双向 LSTM 模型做编码器可能会造成的梯度问题以及训练时间效率问题，本文采用预训练的 BERT 模型作为编码器。BERT 采用了基于 Transformer[21]的双向编码器表示，能够在某一时间步同时获得上下文的编码信息，并且针对不同下游任务仅需微调，不需要进行重新训练。

具体来说，BERT 将预训练任务划分为掩蔽语言模型（Masked Language Modeling，MLM）和下一句预测（Next Sentence Prediction，NSP）两个模块。BERT 能够不受单项语言模型限制，对上下文双向编码的原因是，MLM 起了作用。具体来说，MLM 以 15% 的概率用掩蔽词元（masktoken，[mask]）随机对每个输入序列中的词元进行替换，之后通过训练对被掩蔽的词元进行预测，这样就能得到更深层次的句子的双向表示。NSP 的目的在于理解和学习句子间关系，选择部分句子对 X 和 Y，其中 50% 的 Y 是 X 的下一句，而剩下 50% 则是从语料库中随机抽取的句子。通过学习，判断任一句子对（X，Y）的 isNext 值是真或假，来建模句子对之间的逻辑关系。

由于汉语不存在天然的分词标志，使用 google 官方发布的 BERT 效果并不好，因此本文采用的统一是专为汉语训练和定制的 BERT 版本。

BERT-wwm 是 Cui 等为解决传统 BERT 中以字为粒度的分词，引入了全词掩蔽（Whole Word Masking，WWM）技术而提出的新模型，全词遮蔽指，若某个汉字被遮蔽，则对组成同一词的其余全部汉字进行遮蔽。例如，对于语句：“使用语言模型来预测下一词的可能性”，经过原始的掩蔽处理结果为：“使用语言[MASK]型来[MASK]测下一个词的可[MASK]性”，使用全词掩蔽技术后的结果为：“使用语言[MASK][MASK]来[MASK][MASK]下一个词的[MASK][MASK][MASK]”。该方法改善了模型效果。

本文采用的预训练模型包括 BERT-wwm，RoBERTa-wwm 和 BERT 等 3 种模型。

3.2 主实体增强模块

在开始阶段，将每个输入序列利用分词工具转换为词元。然后，将分词完成的词元序列输入编码器，同时保留序列的相

对顺序。之后对输出的编码进行对齐，将不同长度的编码表示向量序列对齐为同一长度，保留每个词元的相对坐标信息。此后将其转化为矩阵，完成对输入序列的编码。由于相对位置未发生改变，矩阵每个元素仍然保留着与其前词元的位置映射关系，这样就实现了从词元到编码表示向量的映射关系。

更形式化的表示为，对于任一序列 $X_i=x_{i,1},x_{i,2},\cdots,x_{i,m}$，有如下映射关系。

$$f(X_i)=T_i \tag{1}$$

其中，$T_i=t_{i,1},t_{i,2},\cdots,t_{i,n}$ 表示分词后的序列，f 表示映射函数。则对于序列集 $Y=X_1,X_2,\cdots,X_w$，有：

$$f(Y)=T \tag{2}$$

$$q(T)=\boldsymbol{Embedding} \tag{3}$$

其中，$\boldsymbol{Embedding}$ 表示编码表示向量，q 表示映射函数。

同时，有：

$$\forall E_i\in Embedding, g(E_i)=t_{a,b} \tag{4}$$

其中，E_i 为任意编码表示向量，$t_{a,b}$ 为与之对应的词元。g 为映射函数，能够将第 i 个位置的编码表示向量转化为第 a,b 位置的词元。

在完成主实体识别任务后，将所有识别出的实体输入到该模块。该模块会依照主实体的输入顺序生成一个哈希表。索引是输入的相对顺序，值是识别到的全体主实体的编码表示向量。映射关系如式(5)所示：

$$\phi(i)=Entity(i) \tag{5}$$

其中，$Entity(i)$ 表示第 i 个输入的实体。

由于在医学领域，主实体只在“疾病”类别中出现，因此通过将其与任务开始时制作的疾病列表进行比对，将非疾病实体的索引指向空值。此后，遍历哈希表，将所有非空的值加入队列，即为目标主实体集。

3.3 条件层归一化

Wei 等在 CASREL 框架中提出的，将识别出的主实体加入到模型参数中辅助客实体关系识别的方法是，直接将 BERT 输出的编码表示向量和主实体信息简单相加。但直接相加可能会导致梯度爆炸或梯度消失等问题，如果引入归一化方法，则可以使得模型的 loss 变得更加光滑，使得训练更平稳地进行[22]。如果采用批归一化(Batch Normalization，BN)，需要一个相对稳定的均值和方差，而由于 BERT 内部的 transformer 架构，在不同时间步有着各自的均值和方差，这些值相差较大，使用 BN 会导致归一化效果很差，失去了归一化的意义，因此引入层归一化(Layer Normalization，LN)作为归一化的方案。与 BN 不同的是，LN 是对一个中间层的所有神经元做归一化，以 μ 为该层的均值，以 σ 为该层的标准差，则该层的层归一化计算方法如式(6)－式(8)所示：

$$\mu=\frac{1}{H}\sum_{i=1}^{H}\alpha_i \tag{6}$$

$$\sigma=\sqrt{\frac{1}{H}\sum_{i=1}^{H}(\alpha_i-\mu)^2} \tag{7}$$

$$x=relu\left(\frac{\gamma}{\sigma}(\alpha_i-\mu)+\beta\right) \tag{8}$$

其中，H 表示所有节点的数目，α 表示某节点激活前的值，γ 和 β 分别表示模型可训练的缩放和平移向量，x 表示经归一化后的输出。

Su 指出[23]，对于已经预训练好的模型，存在现成的、无条件的 γ 和 β，可以将输入参数转化为与 γ 和 β 同维度的向量，加到训练好的 γ 和 β 中，通过控制输入参数，就可以控制 γ 和 β，从而实现通过条件控制生成的行为。这与本文模型中将主实体信息当作输入参数加入到 BERT 输出的编码表示向量中用于识别客实体的思路不谋而合。Su 用情感分类任务来验证该思路的可行性，取得了较好的结果。本文将是否为客实体当作正面或负面情感，这样就可以将客实体的识别任务转化为一个情感分类任务。若是客实体，则可以在加快识别客实体、关系的速度的同时，提高识别的准确率。通过将上一步得到的主实体序列转化为与 γ 和 β 同维度向量，可以将条件层归一化引入到模型中，具体实现如式(9)－式(11)所示：

$$\gamma'=w_\gamma\cdot s+\gamma \tag{9}$$

$$\beta'=w_\beta\cdot s+\beta \tag{10}$$

$$x'=relu\left(\frac{\gamma'}{\sigma}(\alpha_i-\mu)+\beta'\right) \tag{11}$$

其中，s 表示作为输入信息传入的主实体，w_γ 和 w_β 表示将主实体维度转化为与 γ 和 β 同维度向量的映射矩阵。

4 实验和结论

4.1 数据集

本文采用中国健康信息处理会议 2020 年的共享评测任务 2 所提供的数据集[24]：中文医学信息抽取数据集(Chinese Medical Information Extraction Dataset，CMeIE)。该数据集由郑州大学自然语言处理实验室牵头，联合北京大学计算语言学教育部重点实验室、哈尔滨工业大学(深圳)以及鹏程实验室共同构建。

评测任务的数据选取《儿科学》《临床儿科学》和《临床实践》等医学教材作为注释语料库，共包括 28008 条经过人工标注实体关系的中文医学文本、85 282 个三元组以及预先定义好的 53 种实体关系标签(包含 11 种医学实体类别和 44 种关系类别)组成。该数据集将数据划分为训练集、验证集、测试集 1 和测试集 2 这 4 个部分，具体情况如图 3 所示。

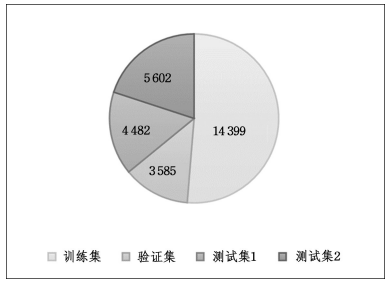

图 3 CMeIE 数据集的数据划分情况

Fig. 3 Data division of CMeIE dataset

在 2020 年组织的基于该数据集的公开测评活动中，最好的成绩由零氪科技基于层叠指针网络和多头选择＋线性关系分类方法取得，在该测试集上其方法 F1 值达到 64.86%，这个成绩比第二名高 1.1%。

4.2 评测指标

本实验采用精确率(Precision，P)、召回率(Recall，R)和 F1 值(F1-Measure)作为该数据集的测评指标。

对于二分类问题，可以将分类结果划分为如下 4 种情况。真正例(True Positive，TP)：实际为真的结果被判断为真。假正例(False Positive，FP)：实际为假的结果被判断为真。真反例(True Negative，TN)：实际为假的结果被判断为假。假反例(False Negative，FN)：实际为真的结果被判断为假。根据这 4 种结果，可以定义精确率 P 和召回率 R。

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R=\frac{TP}{TP+FN} \tag{13}$$

F1 值定义为 P 和 R 的调和平均值：

$$F1=\frac{2\times P\times R}{P+R} \tag{14}$$

4.3 基线模型

本文选择的基线模型包括：1）多头选择模型（Multihead Selection，MHS）[25]；2）以多种不同预训练 BERT 作为编码器的层叠指针网络框架；3）基于 Biaffine 注意力的多头选择模型（Biaffine Based Multihead Selection，BMHS）[26]。

其中，多头选择模型使用 word2vec 技术将单词映射为编码表示向量，使用双向 LSTM 作为编码层，使用 CRF 完成命名实体识别任务，然后将关系抽取任务建模为多头选择问题。该方法将一个序列 w 和关系标签集合 R 作为输入，目标是对于每个词元 $w_i\in w$，标识出最有可能的头向量 $\hat{\boldsymbol{y}}_i$ 以及对应的关系标签 $\hat{r}_i$。

层叠指针网络前文已经介绍，本节不再赘述。

基于 Biaffine 的多头选择模型的主要创新点在于将 Biaffine 注意力机制引入到命名实体识别任务中。其特点是将编码层输出结果 h_n 经过两个前馈神经网络并加入线性偏差，构造头向量及尾向量，然后通过 Biaffine 分类器将对应的头向量和尾向量拼接为实体。

4.4 主实验

本节共设计 6 组实验，其中前 5 组为基准实验，分别使用 mhs，Bmhs，CASREL-BERT，CASREL-BERT-wwm，CASREL-RoBERTa-wwm，SE-CAS 模型在数据集 CMeIE 上进行多轮实验，选取最好的数据作为实验结果。其中 CASREL-X 表示使用了不同的预训练模型作为该模型的编码器。在预先实验中注意到，RoBERTa-wwm 模型的效果表现最好，可能是因为其对模型训练时间更长，批次更大，应用该数据更多，所以本文模型所采用的预训练模型为 RoBERTa-wwm。实验结果均保留 4 位有效数字。实验结果如表 1 所列。

表 1 主实验结果

Table 1 Main experiment results

（%）

模型	准确率	召回率	F1
mhs	52.91	48.23	50.46
Bmhs	62.38	61.61	61.99
CASREL-BERT	61.02	56.28	58.56
CASREL-BERT-wwm	62.57	56.52	59.39
CASREL-RoBERTa-wwm	62.46	57.39	59.82
SE-CAS	**68.27**	**63.03**	**65.55**

可以注意到，本文模型在精确率、召回率和 F1 值各个方面均有较为显著的提升，最终 F1 值达到了最高的 65.55%。其中在精确率方面提升最为显著，相对于基准实验中表现最好的 CASREL-RoBERTa-wwm 框架，提升幅度也为 5.81%，可见通过引入主实体增强模块，减少模型对错误主实体解码的发生，在一定程度上减少了错误关系三元组的产生，提高了模型的精确率。在召回率方面的提升也比较优秀，相对基准实验中表现最好的 Bmhs 模型提高了 1.42%，推测是因为引入了条件层归一化，使得分类任务变得更加明确，对客实体和关系的识别率得到进一步上升，减少了漏查现象的发生。这个数据同样也超过了 2020 年测评时的最好成绩，即64.86%，这足以证明，通过改进层叠指针网络在医学领域起到了较好的应用效果。特别地，由于测评任务中的方法并没有给出具体的实现，其方法细节不易被了解，因此未被作为基线模型对比。

4.5 消融实验

本文针对层叠指针网络提出了两个改进点，为了验证改进的有效性，因此设计了以下 4 组实验作为消融实验：1）未加入两个改进点的原 CASREL 框架在 CMeIE 数据集上的运行结果；2）仅加入了条件层归一化的 CASREL-CLN 模型在数据集上的运行结果；3）仅加入了主实体增强模块的 CASREL-SE 模型在数据集上的运行结果；4）本文提出的新模型 SE-CAS 在数据集上的运行结果。以上 4 组实验所使用的预训练模型皆为 RoBERTa-wwm。其中第一组和第四组实验结果来自 4.5 节。实验结果均保留 4 位有效数字，实验结果如表 2 所列。

表 2 消融实验结果

Table 2 Ablation study results

（%）

模型	准确率	召回率	F1
CASREL	62.46	57.39	59.82
CASREL-CLN	62.68	59.67	61.14
CASREL-SE	64.29	58.35	61.18
SE-CAS	**68.27**	**63.03**	**65.55**

由表 3 分析可知，加入了主实体增强模块后，模型的精确率、召回率相对原模型均有上升，其中精确率上升较为显著，说明主实体增强模块的存在，减小了其他实体的误识别概率，进而提高了识别的精确度，例如对于语句“慢性胰腺炎腹部超声/CT 扫描发现主胰管直径 2～4 mm，或腺体较正常 1～2 倍增大。”，若将“腹部超声”识别为主实体，在客实体关系模块则一定会解码出错误的关系三元组，而主实体增强模块则会直接删除“腹部超声”，进而从根本上避免了这样的错误。

加入条件层归一化后，通过将客实体关系识别转化为分类任务，降低了识别难度，原来的部分因复用主实体效果不好而未被识别出的客实体被成功识别，召回率有了明显的提升。这与之前实验的分析结果一致。

本文提出的新模型在 3 个指标上均显著高于原模型和两个消融模型，SE-CAS 与 CASREL-CLN 的对比可以看出，主实体增强模块的引入同时提高了精确率和召回率，这与先前所得到的结论可以相互印证；与 CASREL-SE 对比可以看到，模型的精确率和召回率得到了大幅提升，说明由于主实体增强模块的存在，减少了识别错误的主实体，使得条件层归一化模块的输入数据正确性提高，减少了因不正确的输入导致的对模型的“反向训练”。

结束语 本文提出了一种面向中文医学文本的主语增强型层叠指针网络模型，改进了原模型中存在的部分缺点，并且将原模型引入到了生物医学领域。经过实验证实，在各项性能指标上均超过了原有的模型和部分其他在生物医学领域实体关系抽取中常用的模型。然而，由于医学伦理等各种问题，目前能够使用的数据集仍然不足，因此训练的结果不能进一步提升。未来，如果出现更大规模的医学文本数据集，该模型将结合新数据集进行进一步训练，从而得到更好的结果。同时，由于层叠指针网络框架在客实体-关系解码模块没有纠错机制，即使在该模块发现了上一步中的错误，也无法对其进行

纠正。后续,若能将纠错机制加入到该模块,则能进一步提高该模型的精确率,拓展该模型的应用范围,提高该模型的应用前景,减轻医务人员的工作负担。

参 考 文 献

[1] WEI Z P, SU J L, WANG Y, et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 1475-1488.

[2] ZAN H Y, GUAN T F, ZHANG K L, et al. Review of entity relation extraction for medical texts[J]. J. Zhengzhou Univ. (Nat. Sci. Ed), 2020, 52(4): 1-15.

[3] ZHENG S C, WANG F, BAO H Y, et al. Joint Extraction of Entities and RelationsBased on a Novel Tagging Scheme[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1227-1236.

[4] LI H, LIU Y J, XIE Q, et al. Distant Supervision Relation Extraction Model Based on Multi-level Attention Mechanism[J]. Computer Science, 2019, 46(10): 252-257.

[5] ZENG X R, ZENG D J, HE S Z, et al. Extracting Relational Facts by an End-to-End Neural Model with Copy Mechanism [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 506-514.

[6] FU T J, LI P H, MA W Y. GraphRel: Modeling Text as Relational Graphs for Joint Entity and Relation Extraction[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1409-1418.

[7] ZENG X R, HE S S, ZENG D J, et al. Learning the Extraction Order of Multiple Relational Facts in a Sentence with Reinforcement Learning[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 367-377.

[8] ZENG D J, ZHANG R H, LIU Q Y. CopyMTL: Copy Mechanism for Joint Extraction of Entities and Relations with Multi-Task Learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 9507-9514.

[9] 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: 4171-4186.

[10] WANG Y C, YU B W, ZHANG Y Y, et al. TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking[C]// Proceedings of the 28th International Conference on Computational Linguistics. 2020: 1572-1582.

[11] MA L B, REN H M, ZHANG X L. Effective Cascade Dual-Decoder Model for Joint Entity and Relation Extraction[DB/OL]. (2021-06-27) [2023-08-22]. https://arxiv. orgabs/2106.14163.

[12] ZHU X B, ZHOU G, CHEN J, et al. Single-stage Joint Entity Relation Extraction Method Based on Enhanced Sequence Annotation Strategy[J]. Computer Science. 2023, 50(8): 184-192.

[13] UZUNER Ö, SOUTH B R, SHEN S Y, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556.

[14] WEI C H, PENG Y F, LEAMAN R, et al. Overview of the BioCreative V Chemical Disease Relation (CDR) Task[C]// Proceeding of the 5th BioCreative Challenage Evaluation Workshop. 2015: 154-156.

[15] SAHU S K, ANAND A, ORUGANTY K, et al. Relation extraction from clinical texts using domain invariant convolutional neural network[C]// Proceeding of the 15th Workshop on Biomedical Natural Language Processing. 2016: 206-215.

[16] ZHANG Y J, LIN H F, YANG Z H, et al. A hybrid model based on neural networks for biomedical relation extraction[J]. Journal of Biomedical Informatics, 2018, 81: 83-92.

[17] LIU Z J, YANG M, WANG X L, et al. Entity recognition from clinical texts via recurrent neural network[J]. BMC Medical Informatics and Decision Making, 2017, 17(2): 53-61.

[18] LUO L, YANG Z H, CAO M Y, et al. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature[J]. Journal of Biomedical Informatics, 2020, 103: 103384.

[19] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234-1240.

[20] ZHANG S H, U S D, JIA Z, et al. Medical Entity Relation Extraction Based on Deep Network and Self-attention Mechanism [J]. Computer Science, 2021, 48(10): 77-84.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[22] SANTURKAR S, TSIPRAS D, ILYAS A, et al. How Does Batch Normalization Help Optimization? [DB/OL]. (2019-04-15) [2023-08-22]. https://arxiv. org/abs/1805. 11604.

[23] SU J L. Conditional text generation Based on Conditional Layer Normalization[EP/OL]. (2019-12-14) [2023-08-22]. https://spaces. ac. cn/archives/7124.

[24] GAN Z F, ZAN H Y, GUAN T F, et al. Overview of CHIP 2020 Shared Task 2: Entity and Relation Extraction in Chinese Medical Text[J]. Journal of Chinese Information Processing, 2022, 36(6): 101-108.

[25] BEKOULIS G, DELEU G, DEMEESTER T, et al. Joint entity recognition and relation ex-traction as a multi-head selection problem[J]. Expert Systems with Applications, 2018, 114: 34-45.

[26] YU J T, BOHNET B, POESIO M. Named Entity Recognition as Dependency Parsing[DB/OL]. (2020-06-13) [2023-08-22]. https://arxiv. org/abs/2005. 07150.

JIANG Zhihan, born in 2003, undergraduate. His main research interests include natural language processing and mathematical optimization.

ZAN Hongying, born in 1966, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. E20-0008671S). Her main research interests include natural language processing and affective computing.