

本科生毕业设计（学术论文）



题 目  **面向癫痫医学文本的命名实体识别**

**方法设计与实现**

学 院  **计算机学院**

专 业  **计算机科学与技术**

学生姓名  **申露涵**

学 号  **2018141461275** 年级 **2018**

指导教师  **段磊**

教务处制表

二Ο二二年五月二十三日

**面向癫痫医学文本的命名实体识别**

**方法设计与实现**

计算机科学与技术

学生 申露涵 指导老师 段磊

**[摘要]** 对癫痫医学文本进行命名实体识别有利于癫痫知识图谱的构建和智慧诊疗系统的开发。传统的基于静态词向量的命名实体识别技术识别效果不佳，而预训练模型BERT虽然能产生更为丰富的动态词向量，提高命名实体识别的精确率，但其语料选取为维基百科和书籍语料库，更适用于通用领域，对专业术语较多的癫痫医学领域的实体识别效果较差。而BioBERT采用BERT的训练方法在生物语料库中进行训练，更适合于癫痫医学领域。针对此问题，提出了EPIMT-NER模型。主要基于BioBERT和双向长短期记忆网络以及条件随机场构建所得，并引入词性、组块和词典特征针对性地提高EPIMT-NER模型在癫痫文本数据集上的准确率。首先，EPIMT-NER模型通过BioBERT生成词向量，将其与词性、组块和词典特征向量拼接，共同传入文本特征提取层。通过双向长短期记忆网络对文本信息进行特征提取，得到所有标签相对应的得分。将得分结果作为条件随机场的输入，通过该层对约束条件进行分析，选择最高预测得分的标签序列作为最终的预测结果输出。通过在人工标注的数据集EPILEPSY上进行对比实验，EPIMT-NER模型的平均精确率达到93.98%，平均召回率为92.08%，平均F1为93.02%，结合对比实验结果可知，EPIMT-NER模型有效提高了对癫痫医学文本进行命名实体识别的效果。

**[主题词]** 命名实体识别；BioBERT；词性特征；组块特征

**Design and Implementation of Named Entity Recognition Method for Epilepsy Medical Texts**

Computer Science and Technology

Student: SHEN Luhan Adviser: Prof. DUAN Lei

**[Abstract]** The named entity recognition of epilepsy medical texts is beneficial to construct epilepsy medical knowledge map and develop intelligent diagnosis and treatment system. The traditional named entity recognition technology based on static word vector has poor recognition effect. Although the pre-training BERT model can generate more dynamic word vectors and improve the accuracy rate of named entity recognition, its training prediction is based on Wikipedia and book corpus, and it is more suitable for general field, and the entity recognition effect is poor in epilepsy medical texts with more professional terms. However, BioBERT adopts BERT's training method to conduct training in biological corpus, which is more suitable for epilepsy medical texts. To solve this problem, an EPIMT-NER model was proposed, which was mainly constructed based on BioBERT and BiLSTM as well as Conditional Random Fields. Part of speech, chunking and dictionary features were introduced to improve the accuracy on epileptic text datasets. First of all, EPIMT-NER model generates word vector through BioBERT, and splines it with part of speech ,chunking and dictionary feature vector which is jointly passed into the text feature extraction layer.Feature extraction of text information is carried out through bidirectional long and short-term memory network, and the scores of all tags are obtained. The results are taken as the input of conditional random fields. The constraint conditions are analyzed through this layer, and the tag sequence with the highest predicted scores is selected as the final prediction result output. Through the comparison experiment on the manually annotated data set EPILEPSY, the precision of EPIMT-NER model reaches 93.98%, the recall rate is 92.08%, and the F1 is 93.02%. Combined with the comparative experimental results, it can be verified that EPIMT-NER model can effectively improve the recognition effect of epilepsy medical texts.

**[Key Words]** Named entity recognition; BioBERT; Part-of-speech ; Chunking

**目 录**

[1 绪论 1](#_Toc3819)

[1.1 研究背景及意义 1](#_Toc28152)

[1.2 国内外研究现状 2](#_Toc32479)

[1.2.1 基于规则和词典的方法 2](#_Toc1138)

[1.2.2 基于统计机器学习的方法 2](#_Toc17303)

[1.2.3 基于深度学习的方法 2](#_Toc27777)

[1.3 论文主要工作 3](#_Toc31974)

[1.4 论文组织与结构 3](#_Toc12222)

[2 相关工作 4](#_Toc6860)

[2.1 文本词向量表示模型 4](#_Toc32717)

[2.1.1 Word2Vec 4](#_Toc2941)

[2.1.2 BERT模型 5](#_Toc28866)

[2.2 循环神经网络 6](#_Toc16034)

[2.2.1 循环神经网络 6](#_Toc22503)

[2.2.2 长短期记忆网络 7](#_Toc20376)

[2.2.3 双向循环神经网络 8](#_Toc31095)

[3 EPIMT-NER模型 9](#_Toc30395)

[3.1引言 9](#_Toc17723)

[3.2模型整体架构 9](#_Toc8707)

[3.3嵌入层 10](#_Toc8513)

[3.3.1 BioBERT生成词向量 10](#_Toc22390)

[3.3.2 词性标注和组块分析 10](#_Toc31481)

[3.3.3 词典特征 12](#_Toc22589)

[3.3 双向长短期记忆网络 13](#_Toc23159)

[3.4 条件随机场 13](#_Toc29029)

[4 实验过程与结果分析 16](#_Toc20666)

[4.1数据来源 16](#_Toc3978)

[4.2评估标准 17](#_Toc8479)

[4.3实验设置 17](#_Toc5378)

[4.4 实验对比及结果分析 18](#_Toc18278)

[4.4.1 EPIMT-NER模型实验结果及分析 18](#_Toc11844)

[4.4.2 特征引入对模型性能的影响 19](#_Toc18169)

[4.4.3 词向量对模型性能的影响 20](#_Toc32413)

[4.4.4 探究在公开数据集上模型的性能 21](#_Toc21845)

[5 总结和展望 23](#_Toc3954)

[5.1 总结 23](#_Toc16576)

[5.2 展望 24](#_Toc32409)

[参考文献 25](#_Toc15796)

[声 明 26](#_Toc3228)

[致 谢 27](#_Toc22700)

# 1 绪论

## 1.1 研究背景及意义

癫痫是神经科的一种常见疾病，当大脑神经元突发性异常放电时，大脑会出现短暂的功能障碍，根据不同的癫痫类型出现不同的发作症状。据最新的流行病学资料显示，中国癫痫总体发病率达到了7‰，约有900万患者，已仅次于头痛患者。它不仅致残率、致死率高，对于女性患者的生育能力和后代健康也有着严重的影响。从青春期到妊娠和生产的全过程中，女性癫痫管理相当复杂，且目前已发现国际指南在很多方面不适用于我国患者，如果能从癫痫领域的文献中提取出有价值的信息有利于帮助构建一个癫痫领域的学习型知识图谱，并可在此基础上开发一个女性癫痫智能诊疗系统，辅助临床决策，造福中国癫痫女性。

而这些年科学技术水平快速发展，计算机应用范围不断扩大，各行各业都在产生大量的数据信息。这些使得计算机数据呈现井喷式增长，加大了信息获取的难度。比如，癫痫领域的文献中就蕴含着丰富的信息，但是多数情况下，计算机是无法直接识别并提取出有效信息的。因此，自然语言处理（Natural Language Processing）开始飞速发展，提出了多项技术去解决人与机器之间无法沟通的问题。信息抽取就是NLP的子任务之一，主要是对指定类型的实体、实体之间的语义关系和文件中的事件进行抽取。命名实体识别因此而生，成为NLP的关键性基本任务之一。它的基本定义就是从各类文本中将特定的实体识别出来识别。通常是人名、地名、组织机构名等。但在不同的领域有不同的需求和实体定义，这些实体大多由人类预先定义，如癫痫医学文本中的药物名实体、方法类实体等。这也是关系抽取、事件抽取、知识图谱构建、机器翻译、问答系统等NLP任务的基础，为后续工作的开展奠定基石。因此，通过学习和构建模型提高癫痫医学文本命名实体识别的准确率，可以作为后期构建知识图谱和开发女性癫痫智慧诊疗系统的基础。

命名实体识别一直是自然语言处理领域具有挑战性的热门课题，主要研究目的是提高实体识别的效果。纵观命名实体识别技术的发展历程可以看出随着计算机技术的发展，命名实体识别技术也在不断发展，学者们提出了许多能提高实体识别准确率的方法。最初是根据自定义的词典和规则模板去匹配实体来完成命名实体识别，后来提出了统计机器学习方法，减少对人工的依赖，再到如今应用范围最广的基于深度学习的模型。但由于命名实体本身具有的复杂性、局限性、开放性、多变性等特点，在实际应用中还有许多亟待解决的问题，仍然具有挑战性和研究价值。尤其是在特有领域方面，大部分方法的迁移能力不强。通用研究方法在特定数据集上可能不能发挥好的效果。

因此本文利用深度学习技术，构建面向癫痫医学文本的命名实体识别模型，希望能有效提高对癫痫医学文本上的识别效果。从癫痫文献中提取出有效信息，有利于对信息进行保存分析和继续利用，为后续知识图谱的构建和智慧诊疗系统的开发打下基础，具有较强的社会意义。

## 1.2 国内外研究现状

命名实体识别既是分类问题，又是序列问题，通常被称为序列标注问题，是自然语言应用的基础。1991年，Rau等人[[[1]](#endnote-0)]在第七届IEEE大会上首次发表了相关研究文献，初步提出了命名实体识别的概念。1996年召开MUC-6会议[[[2]](#endnote-1)]，在本次会议中命名实体识别成为信息领域的一个子任务，开始成为很多学者的研究热点。近年来，计算机技术不断发展革新，命名实体识别技术也随之飞速发展。在技术发展的初期阶段，多采用人工定义的词典和规则模板进行识别；后来采用传统机器学习的方法，条件随机场模型应用甚广；近年来深度学习技术的发展进一步提高了命名实体识别的性能，成为目前应用范围最广的技术。目前很多学者将多种方法和模型相结合以减少单个技术的局限性，提高命名实体识别的效果。然而由于命名实体的复杂性、局限性、开放性，在实际应用中还有许多亟待解决的问题，仍然具有挑战性和研究价值。

1.2.1 基于规则和词典的方法

在早期的命名实体识别的研究中多使用本类方法，其主要是通过匹配来完成识别。其中基于词典的方法[[[3]](#endnote-2)]是通过构建实体词典，将文本与词典逐一匹配。这种方法实现较为简单，但对词典的要求较高，识别效果严重依赖于词典的构建。词典内容较少则识别效果不佳，但要构建较为完善的词典对人工的要求较高，耗费大量人力资源，效率较低。基于规则的方法则需要从文本特点入手，定义合适的规则进行匹配。这类方法需要人工定义规则模板，根据任务和数据集的不同，规则模板也要重新改写，整个设计过程繁琐复杂、耗费大量人力和时间资源，准确率严重依赖于人工设计的合理性、准确性。这两种方法都不易在其他数据集或者领域应用，迁移性差，对人工的需求和依赖较多，局限性比较大。

1.2.2 基于统计机器学习的方法

在传统机器学习方法中，常将命名实体识别任务看作序列标注问题，主要通过标注好的文本进行训练，利用训练好的模型进行识别，将不同的字符映射为对应的标签。这种方法的效果严重依赖于特征的选取。最早提出的隐马尔可夫模型[[[4]](#endnote-3)]结构简单，训练效率高，但无法结合上下文信息考虑。最大熵模型[[[5]](#endnote-4)]和支持向量机模型相对隐马尔可夫模型来说正确率更高，但识别和训练的速度较慢，且不适用于训练大规模样本。在命名实体识别任务中，条件随机场[[[6]](#endnote-5)]的应用更为广泛，且有良好的表现，可以利用内部和上下文信息。基于机器学习的方法减少了对人工的依赖，提高了实体识别的准确率。但这类方法对语料库特征选取的依赖较大，同时识别策略单一，所以效果也比较有限。如何改进模型并提高模型的识别性能依然是众多学者研究的核心。

1.2.3 基于深度学习的方法

近年来，随着计算机硬件水平的发展，算力限制逐渐被突破，深度学习技术迎来了新的发展机会。同时，词的分布式表示（word embedding）的提出解决了独热编码不能计算词汇相似度的问题，也不易产生维度爆炸情况，得到了广泛的应用。因此，结合神经网络的命名实体识别技术成为现在研究的热门方向。2019年以前，BiLSTM-CRF模型[[[7]](#endnote-6)]是NER任务的主流模型之一，它采用两个相反方向的LSTM提取上下文信息，再通过条件随机场对结果进行修正，避免输出不合理的标签序列。随着注意力机制的发展，利用注意力机制提高模型并行能力的Transformer[[[8]](#endnote-7)]结构被人提出。随后，谷歌提出的BERT模型采用多头注意力机制，以Transformer为基础实现了双向语义表征。此后，由于对BERT模型[[[9]](#endnote-8)]进行简单微调就能取得不错的效果，BERT开始逐渐成为目前应用最为广泛的模型。相比于传统方法，深度学习方法是一个端到端的训练过程，可以主动学习文本特征，减少对人工的依赖。同时，大量研究表明在命名实体识别任务中，深度学习方法表现更出色，识别效果更好。目前很多学者将多种方法和模型相结合以减少单个技术的局限性，提高命名实体识别的效果。如Wang等人[[[10]](#endnote-9)]将BERT模型与BiLSTM-CRF模型结合进行中文命名实体识别。之后又有很多学者将模型改进应用于其他特定领域，如许力等人[[[11]](#endnote-10)]加入词性分析和组块分析模块将BERT-BiLSTM-CRF模型应用于生物医学文本领域，都取得了不错的效果。

## 1.3 论文主要工作

论文研究的主要内容是面向癫痫医学文本进行命名实体识别任务的方法设计与实现，为了提高模型的精度和识别效果，本文主要工作如下：

* 构建癫痫医学文本数据集EPILEPSY。对癫痫医学文献进行爬取和预处理，按照BIO格式进行人工标注，B表示一个实体的头部，I是实体中间，O是非实体标签。
* 提出并构建了EPIMT-NER（Epilepsy Medical Text Named Entity Recognition）模型，利用BioBERT生成动态词向量，提高模型的精度，加入词性、组块和词典特征，加强对癫痫医学文本信息的针对性，提高模型的识别效果。通过双向长短期记忆网络提取文本特征，最后利用条件随机场对提取结果进行优化输出。
* 在癫痫医学文本数据集EPILEPSY上利用EPIMT-NER模型进行命名实体识别，并设置对比实验，验证了模型的有效性。

## 1.4 论文组织与结构

本论文主体由五个部分组成，各部分作用分别为：

第1章主要介绍选题背景和意义、研究现状、主要工作以及本篇的组成结构。

第2章主要从文本词向量表示模型和循环神经网络两个方面对论文所用到的相关技术理论基础进行阐述，为后面模型构建提供理论依据。

第3章提出了EPIMT-NER模型，从模型的整体架构到每一层的具体设计与实现进行了阐述。

第4章为实验部分，主要介绍了数据集的来源与标注规范、模型评估标准、实验环境与设置，并设置对比实验，验证模型的有效性。

第5章总结了论文的主要工作，记录了实验过程中遇到的问题并给出了解决方案，同时指出不足和未来改进工作。

# 2 相关工作

论文提出了EPIMT-NER模型，该模型使用神经网络对词向量进行特征提取，本章将对相关工作进行介绍。首先对常用的生成词向量的预训练模型Word2Vec和BERT的原理进行阐述。其次，论文使用循环神经网络层对词向量进行文本特征提取，本章将对网络层的基础原理进行解释阐述。

## 2.1 文本词向量表示模型

由于计算机无法直接识别人类语言，机器学习不能直接将文本作为输入进行学习，需要将文本表示为向量形式，也称之为词向量。目前主要有两种词向量表示形式：离散表示和分布式表示。离散表示的代表独热编码（one-hot），它的向量大小即为词表大小，只在单词对应索引位置标1，是最简单的词向量表示方法。但是这种方法表示的词向量过于稀疏且不能表达出文本的距离，在本文中没有使用，仅作简单介绍。分布式表示的模型有很多，论文使用基于BERT改进的BioBERT生成词向量，Word2Vec生成词性向量和组块向量，因此本节将对Word2Vec和BERT的原理作简要介绍。

2.1.1 Word2Vec

Word2vec[[[12]](#endnote-11)]是一种生成静态词向量的工具，它主要认为词和它相邻近的词之间是有联系的，如果两个词的上下文相似，就认为这两个词也是相似的，即可用词来预测词。它将每个词的独热编码表示的向量作为输入，通过训练将稀疏向量映射为稠密词向量。根据输入输出层的不同，分为CBOW模型和Skip-gram模型。其中，CBOW模型是通过输入的上下文来预测中心词，而Skip-gram是依据中心词推测其邻近词。Word2Vec使用了浅层神经网络，训练的速度快，相对于独热码来说能获取到更好的表征。但静态词向量在调用时是固定的，不会根据上下文或者语义的不同产生变化，每个词使用了唯一的词向量，因此Word2Vec没有处理一词多义的能力，在对语义有要求的任务中不太适用。

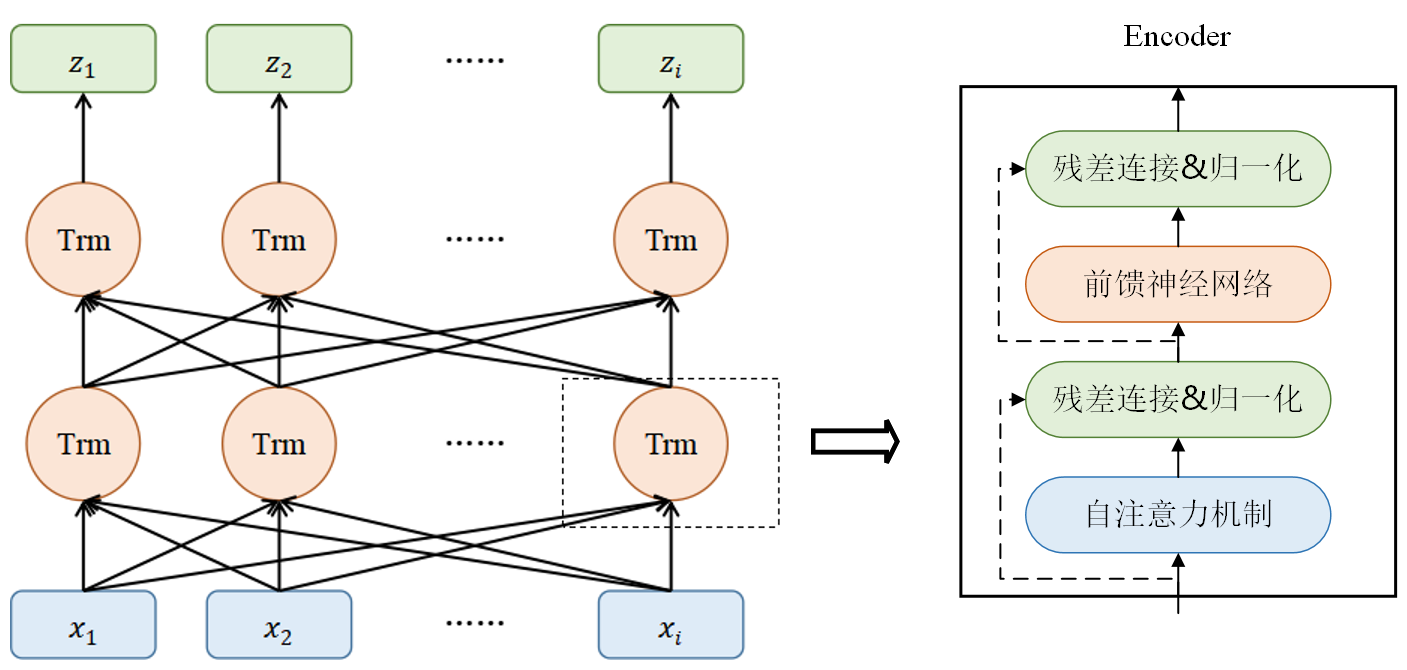


图2.1 BERT结构示意图

2.1.2 BERT模型

BERT是在transformer的基础上衍生出的预训练模型，它利用新的预训练方法masked language model（MLM）在维基百科和书籍语料库中进行预训练，再根据下游任务的不同对模型进行微调。与传统的单项语言模型或者两个单项语言模型简单拼接相比，BERT不受单向语言模型所限制的原因并不是使用两个transformer双向扫描输入序列，这种方法虽然直接有效，但模型参数太多，训练成本很高。它构建了两个新的预训练任务：随机遮蔽词语预测（MLM）和是否下一句（NSP），使得模型可以双向思考，取得更好的效果。MLM方法是以15%的概率用[MASK]随机替换掉训练序列中的token，然后预测出[MASK]位置原有的单词。这种方法能更好地结合上下文信息，生成动态词向量。MLM任务更关注token层面的语义信息，不能直接获取句子间的表征，因此BERT增加了NSP任务，使模型对句子间关系的理解更强，可以获取句子级别的上下文关系。BERT的双向表征能力使得它从被提出开始就得到了广泛的应用，并在一系列句子级和字符级任务中取得了不错的效果，功能十分强大。

BERT的输入层是文本中每个词的原始向量，可以由随机初始化所得，也可以由Word2Vec初步训练所得。主要由token embedding、position embedding、segment embedding相加而成。其中token embedding是单词原始的词向量；由于transformer无法获取单词的位置信息，因此加入绝对位置position embedding来保存单词在句中的序列位置信息；同时模型可能有多个句子的输入，为了更好地适配下游任务，加入segment embedding表示单词归属于哪个句子的信息。BERT的模型主体结构是多层transformer的encoder堆叠而成，如图2.1所示。每一个encoder由两层组成，包含self-attention层和前馈神经网络层。其中自注意力机制能不局限于当前的单词，从而能获取上下文的语义信息。当self-attention处理完数据后会将结果传给前馈神经网络层，前馈神经网络的计算可以并行，将得到的输出传递到下一个encoder。

首先self-attention会将词嵌入X与随机初始化所得的权重矩阵，，相乘，给每个单词计算出三个新的向量值，得到新的向量矩阵：Q（Query），K（Key），V（Value）：

(2.1)

将Q与K作矩阵乘积得到self-attention的分数值score，这个值反映了某个单词对其他单词的关注权重。再将score值作归一化计算和softmax计算，得到每个词对当前位置词的相关性大小。再通过矩阵乘法将需要增加的注意力进一步增大，得到当前节点的值，具体的计算公式如下：

(2.2)

此外，为了提高精度和效率，BERT引入了“multi-headed”attention（多头注意力机制），将*h*个self-attention集成到一起，相关计算公式如下：

(2.3)

(2.4)

BERT主要有pooler output、sequence output两类输出，可以根据下游任务的不同选择不同的输出。在BERT预训练模型的基础上根据不同的下游任务进行微调（fine-tune），可以减小资源成本开销，同时获得更好的性能效果。

**2.2 循环神经网络**

2.2.1 循环神经网络

传统的神经网络模型不能关注到上一个时刻出现的信息，只会关注于当前时刻，对需要根据前文判断后文的情况处理效果不佳。因此希望有一个神经网络可以“记住”已经出现过的信息，循环神经网络（RNN）由此而生。具体结构示意图如图2.2和图2.3所示，结合本文任务和模型而言，神经网络层的输入应为预训练模型生成的特征向量。从图2.2的左边部分可以看出，RNN带有一个指向自身的环，按照时间线展开即如图2.2的右边部分所示，它可以在相邻时刻之间传递信息。其中W是每个时间点的权重矩阵，在整个过程共享一个W。

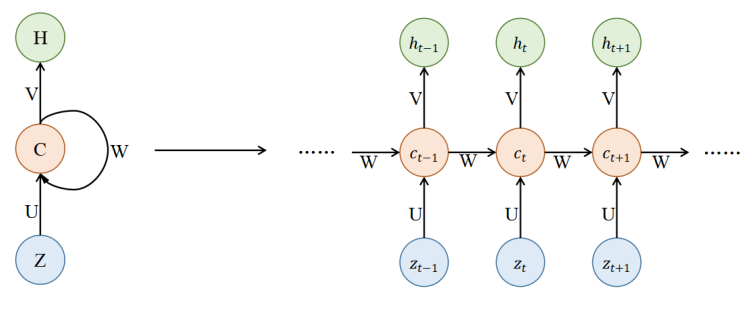


图2.2 循环神经网络结构图

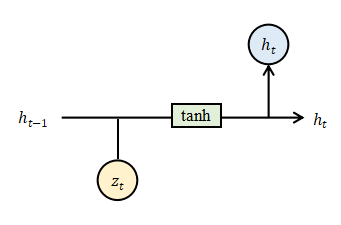


图2.3 RNN单元结构图

2.2.2 长短期记忆网络

LSTM就是循环神经网络的一种，全称是长短期记忆网络。主要解决了RNN梯度消失和爆炸的问题，进而提高捕捉长期依赖关系的能力，其单元结构如图2.4所示。LSTM的输入为：，，，输出为 ，。其中表示t时刻LSTM的单元状态，表示t时刻LSTM的输出值，表示t时刻网络的输入值。LSTM主要通过门控单元来选择性地存储信息，这也是它能处理循环神经网络长期依赖问题的原因。由图2.4所示，LSTM单元结构中主要包含三个门：遗忘门（forget gate），输入门（input gate）和输出门（output gate）。相关计算公式如下：

(2.5)

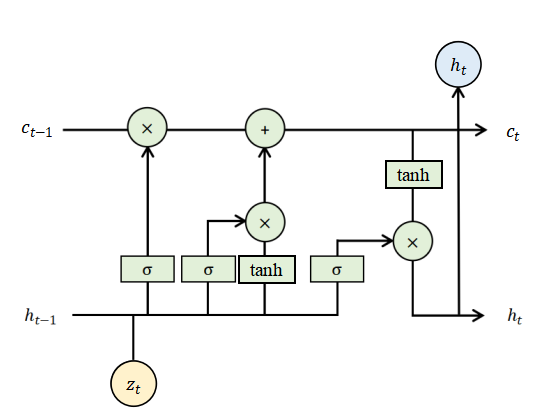
(2.6)

(2.7)

(2.8)

(2.9)

(2.10)

图2.4 LSTM单元结构图

其中为sigmoid激活函数，将输出转化为0到1之间的值，以此作为门控状态。tanh函数表示双曲正切函数，输出值在-1到1之间。w表示门控单元的权重矩阵。b是对应门控单元的偏置向量。

遗忘门可以对上一时刻的信息选择性地过滤，忘记不重要的，只保留重要信息至当前时刻；输入门会对当前的输入有选择的记忆，和tanh函数共同控制信息的更新；输出门决定当前单元状态保留多少信息作为输出。

2.2.3 双向循环神经网络

根据2.2.1节可知RNN具有记忆功能，可以结合上文信息综合考虑。但是它只保留了之前的信息，并不能对下文信息进行处理。要想获取下文信息，可考虑再使用一个相反方向的循环神经网络实现，因此对循环神经网络进行改进。在循环神经网络的基础上，增加了一个反向运算，可以理解为将输入序列反转，对反转后的序列做一次RNN计算，即得到了反向RNN结果。最后把两个不同方向上的RNN结果进行拼接，就得到了最后的输出结果。这种神经网络被称为双向循环神经网络（BiRNN）。

# 3 EPIMT-NER模型

## 3.1引言

目前在命名实体识别任务中还没有针对癫痫领域的公开数据集，本文对癫痫领域公开文献进行爬取和数据预处理，作为实验数据集，利用深度学习技术进行文本命名实体识别，为后续参与开发女性癫痫智慧诊疗系统提供基础工具并打下坚实的学习基础。

针对癫痫医学文本中医学专业词汇量大，数据集相对较少的特点，本文采用EPIMT-NER模型对癫痫文本进行命名实体识别。任务的目标是通过EPIMT-NER模型在癫痫医学文本数据集上实现命名实体识别，并将识别结果与现有的其他方法进行对比，验证模型的有效性。

## 3.2模型整体架构

EPIMT-NER模型的整体架构如图3.1所示，由输入层、嵌入层（embedding）、文本特征提取层、条件随机场优化层和输出层组成。嵌入层主要由BioBERT预训练模型生成的词向量与Word2Vec对应训练的词性标注向量、组块标注向量和词典特征向量进行拼接（contact）所得，结果将作为文本特征提取层的输入，通过BiLSTM模型对文本信息进行特征提取，得到每个单词对所有标签相应的得分，将特征提取层的输出作为条件随机场的输入，通过CRF模型对约束条件进行分析，选择最高预测得分的标签序列作为最终的预测结果输出。

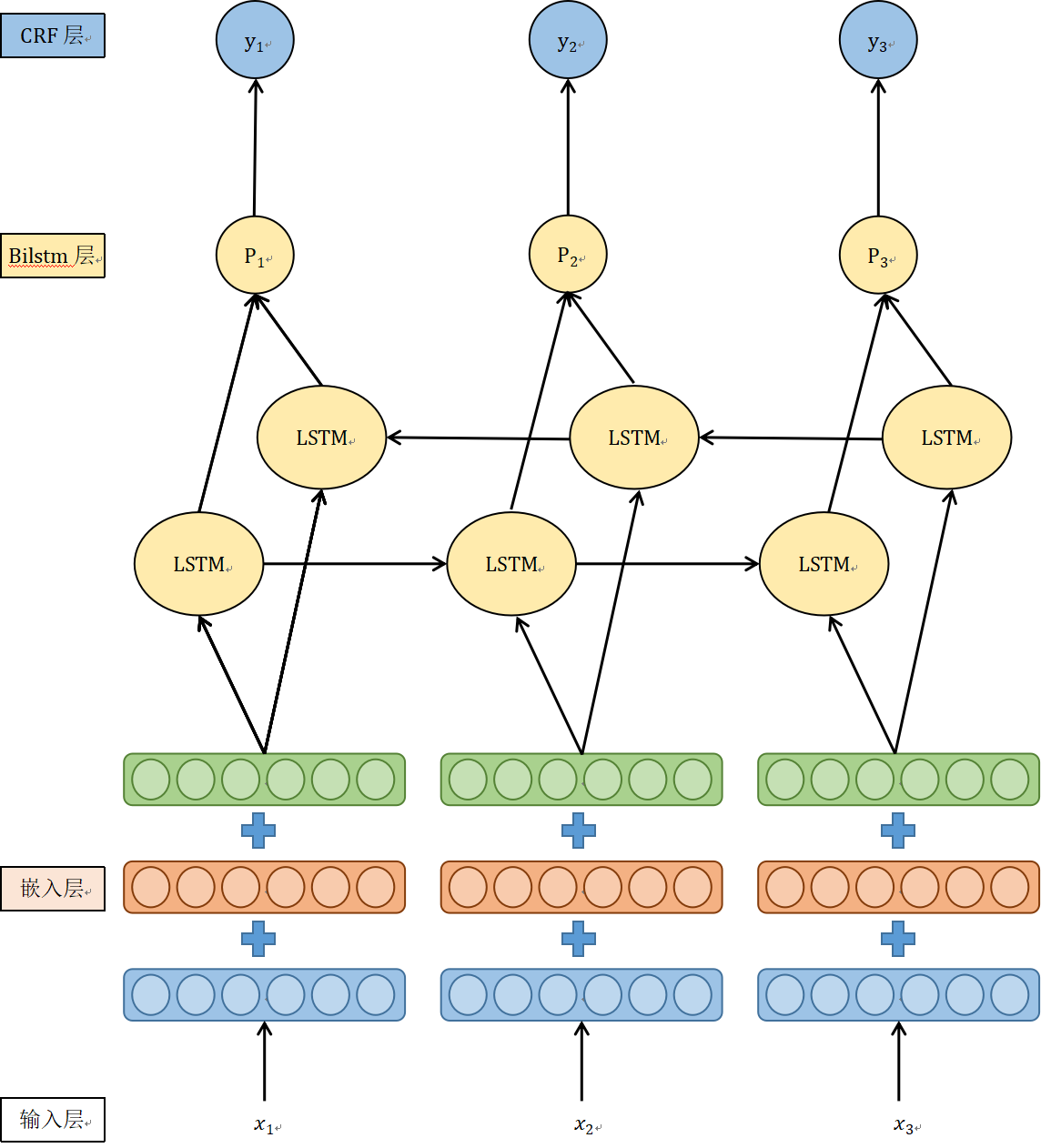


图3.1 模型整体架构图

## 3.3嵌入层

3.3.1 BioBERT生成词向量

模型的输入是数据集经过预处理后得到序列标注文本，但其无法被直接识别和计算，需要转化为计算机可以理解的形式。因此，在本文命名实体识别的任务中，首先要将输入文本转化为词向量，进而才能将其输入文本特征提取层。

在2.2节中对预训练模型进行了简单的介绍，并学习了两种预训练模型：Word2Vec和BERT。Word2Vec无法解决一词多义的问题，相较之下BERT的功能更为强大，能在使用较少数据标注的情况下生成更丰富的动态词向量。而在BERT之后又提出了很多基于BERT的预训练模型，在保留BERT优势的情况下对模型进一步改进，如BioBERT。

BERT预训练模型主要是通过维基百科和书籍语料库的数据进行的预训练，在通用领域中识别效果比较好，而针对特有领域数据时存在一定的局限性，可能会使得命名实体识别的效果并不出色。尤其是近年来随着生物医学文献数量的增加，学者们也更关注如何从生物医学文本中识别出有效信息。在生物医学领域，由于专业词汇的数量较大，普通的文本挖掘方法并不能很好的识别出专业用语，所以Lee 等[[[13]](#endnote-12)]学者提出了BioBERT模型，极大的提高了生物医学文本挖掘的性能，在命名实体识别、关系抽取和生物问答等任务中都取得了出色的表现。

BioBERT是针对生物医学领域的预训练模型，与BERT有着相同的结构。它在BERT模型的基础上针对生物医学领域做了专业的数据收集与训练。主要方法如下：首先，通过BERT的权重初始化BioBERT。然后，使用生物医学领域的语料库（PubMed摘要和PMC全文文章）对BioBERT进行预训练。因此，同样可以在BioBERT预训练模型的基础上进行任务微调，以最小的成本获得更好的性能效果。

本文的主要任务是针对癫痫医学文本做命名实体识别，结合上文所述，BioBERT具有如下优势：

* 具有与BERT一样的双向语言表征，可以结合上下文信息考量，相较于Word2Vec解决了一词多义的问题，能获得更丰富的动态词向量。
* 模型使用了新的训练方法MLM和NSP，使之可以获得句子级别的语义表征。
* BioBERT的输入输出层可适配多项下游任务，且迁移性好。
* 采用BioBERT预训练后微调的成本低，资源开销小。
* BioBERT使用了生物医学领域的语料进行预训练，对本文的癫痫医学文本有着非常好的适用性。

因此本文选择BioBERT模型作为预训练模型，获取动态词向量。

3.3.2 词性标注和组块分析

在自然语言处理任务中，词性标注通常是作为预处理的步骤，主要是指将语料库中的每个单词根据其含义和上下文确定词性的过程。比如名词、动词、形容词、副词等，如表3.1所示为英文词性标注对照表。大量研究表明，词性标注和命名实体识别任务具有强相关性，可以辅助判断实体位置，提高精确率。如在对癫痫文本数据进行命名实体识别时，通过对数据进行分析，我们其实可以提取出一些明显的特征，将这些明显的特征与模型相结合，可以提高识别的精度。在癫痫文本中大部分实体的词性都是名词词性，比如药物类实体medication，病因分析中的致病源实体etiology等，而动词、副词等可以明显判定为非实体。这种特征对我们改进模型、提高识别准确率有极大的帮助。因此，我们可以将文本进行词性标注，再生成词性向量，作为特征与词向量一起输入文本特征提取层，辅助命名实体识别。

命名实体识别的难点之一是实体边界的确定。对于给定的一个句子序列，单词之间可能存在内部联系，比如一些名词块、动词块等。组块分析可以识别出句子中各种互不交叉、没有嵌套的句块，帮助模型更好的识别实体边界。通常来说词性标注的结果对于组块分析而言非常有价值，可以在其基础上进行组块分析。在本文的任务中，增加组块分析特征能辅助确定实体边界，有效提高实体识别的精确率。

表3.1词性标注对照表

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 代码 | 词性 | 代码 | 词性 | 代码 | 词性 | 代码 | 词性 |
| CC | Coordinating conjunction | LS | List item marker | PRP$ | Possessive pronoun | VBD | Verb, past tense |
| CD | Cardinal number | MD | Modal | RB | Adverb | VBG | Verb, gerund or present participle |
| DT | Determiner | NN | Noun, singular or mass | RBR | Adverb, comparative | VBN | Verb, past participle |
| EX | Existential there | NNS | Noun, plural | RBS | Adverb, superlative | VBP | Verb, non-3rd person singular present |
| FW | Foreign word | NNP | Proper noun, singular | RP | Particle | VBZ | Verb, 3rd person singular present |
| IN | Preposition or subordinating conjunction | NNPS | Proper noun, plural | SYM | Symbol | WDT | Wh-determiner |
| JJ | Adjective | PDT | Predeterminer | TO | to | WP | Wh-pronoun |
| JJR | Adjective, comparative | POS | Possessive ending | UH | Interjection | WP$ | Possessive wh-pronoun |
| JJS | Adjective, superlative | PRP | Personal pronoun | VB | Verb, base form | WRB | Wh-adverb |

综上所述，针对癫痫医学文本数据的特征，EPIMT-NER模型引入了词性标注和组块分析标注共同作为文本的特征进行学习。GENIA Tagger是一个针对于生物语料的词性标注工具，可以通过该工具获得词性标注和组块分析标注。使用该工具对癫痫医学文本数据集EPILEPSY进行标注后的部分标注示例如表3.2所示。在构建完词性标注和组块分析标注后，将其转换为词向量表示，本文采用2.1节中提到的Word2Vec工具生成词性向量和组块向量。

最终输入层的词向量由BioBERT生成的词向量与Word2Vec生成的特征标注向量拼接而得。

表3.2 词性标注和组块分析标注生成示例表

|  |  |  |
| --- | --- | --- |
| 单词 | 词性 | 组块 |
| Ask | VB | B-VP |
| the | DT | B-NP |
| patient | NN | I-NP |
| or | CC | I-NP |
| caregivers | NNS | I-NP |
| to | TO | B-VP |
| confirm | VB | I-VP |
| that | IN | B-SBAR |
| this | DT | B-NP |
| is | VBZ | B-VP |
| a | DT | B-NP |
| habitual | JJ | I-NP |
| seizure | NN | I-NP |

3.3.3 词典特征

在癫痫医学文本中，部分实体的边界是很难确定的，如autistic spectrum disorder（自闭症障碍）、magnetic resonance spectroscopy（核磁共振光谱学）、childhood absence epilepsy（儿童失神癫痫）等，这类实体在通用领域数据中不常见，大多为与癫痫医学领域相关的专有名词。同时，一个词单独出现时的实体类型与其嵌套出现的实体类型有可能是不同的。比如epilepsy单独出现时大多属于epilepsy实体，标记为B-epilepsy。当它在词组childhood absence epilepsy中出现时属于classification实体，应该标记为I-classification。对上述特定实体而言，单纯使用组块分析来确定实体边界，效果往往是不好的。虽然基于词典进行识别的方法现在已经不常使用，但词典信息有利于EPIMT-NER模型进行特定实体边界的确定。因此，本文通过对华西专家提供的实体词典进行词典特征的提取，增强语义信息，针对性地提高癫痫医学文本命名实体识别的效果。双向最大匹配法通常用于中文分词，在提取词典特征时可以将英文按照空格划分，每个单词就类似于中文分词中的一个字。借助实体词典，采用双向最大匹配法就能提取词典特征。双向最大匹配法的核心是比较正向和逆向的最大匹配结果，选择词组数量少的划分结果以获得最大匹配实体。当词组数量相同时，选择独立成词的词组的数量更少的结果。对于输入的句子序列，如果匹配成功就标记为1，其他标记为0。以输入序列“Conditions such as syncope associated with cardiac arrhythmia”为例，借助实体词典标记词典特征如表3.3所示：

表3.3 词典特征示例

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 输入 | Conditions | such | as | syncope | associated | with | cardiac | arrhythmia |
| 标签 | O | O | O | B-differential | O | O | B-comorbiditity | I-comorbiditity |
| 词典 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |

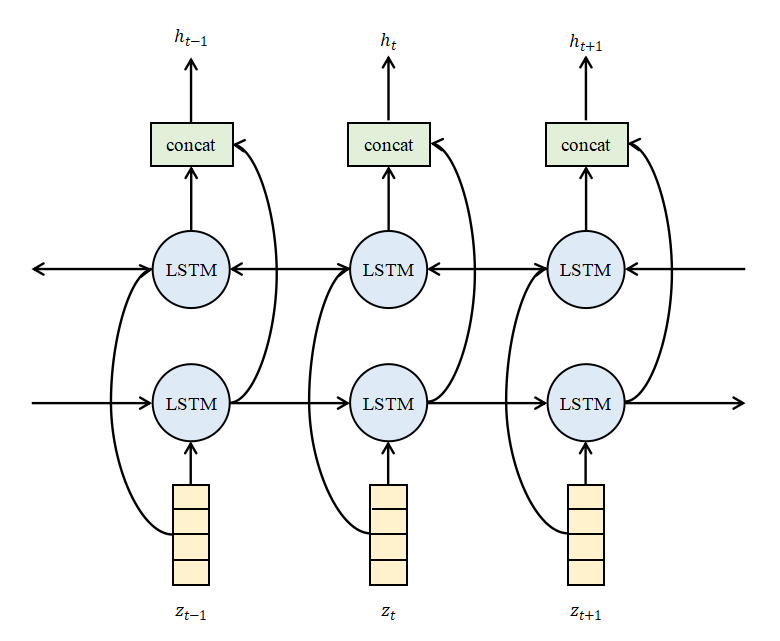


图3.2 双向长短期记忆网络结构图

## 3.3 双向长短期记忆网络

本层为文本特征提取层，将在3.2节中获得的词向量作为本层的输入进行特征提取。2.3节中介绍了双向循环神经网络和长短期记忆网络，前者可以结合上下文信息考虑，但也同样存在循环神经网络中长期依赖的问题，而后者可以解决这类问题，但只能获取上文信息。因此综合考虑，长短期记忆网络（BiLSTM）结合了二者的优势，结构示意图如图3.2所示。它是由一个前向的LSTM和一个后向的LSTM结合而成，可以更好的捕捉双向语义依赖，获得上下文信息。输入序列分别通过前向LSTM和后向LSTM得到两个方向上的向量和,最终结果由和,拼接而得，即：

（3.1）

BiLSTM的输入输出维度保持一致，对于序列，通过双向循环神经网络层得到，包含了每个词分别对应于每个标签的概率。理论上讲，可以将此时输出的标签概率序列进行一个max求值，直接输出单词对应的概率最大的标签。但在实际应用中，这也可能会出现不合逻辑的标签，因此引入了条件随机场CRF进行条件约束，将本层的输出作为输入给到CRF层。

## 3.4 条件随机场

2001年，Lafferty等首次在论文中给出条件随机场（CRF）的概念，主要用于标注和分析序列数据。条件随机场模型是一个判别式模型，通常是在给定输入序列的条件下，对输出序列的条件概率建模。相关定义具体如下：

设，，当满足：

（3.2）

对任意顶点m成立时，则称为条件随机场。

通常使用的是线性链条件随机场，本文针对癫痫医学文本做命名实体识别任务时，输入序列和输出序列满足：

（3.3）

符合线性链条件随机场的定义，其结构示意图如图3.3所示。

结合图举一个例子简单说明一下CRF的概念：设输入序列为“我喜欢四川”，采用BIO格式对句子进行标注，即输出标注序列中的取值都为集合，有多种随机组合的方式。确定字符标签后，输出标注结果为“我O喜O欢O四B-LOC川I-LOC”，这就形成了一个随机场。由于每个字都有“B”、“I”、“O”三种标注选择，最终的标注序列也会存在多种可能，如（O,O,I,B,O,）序列。但I是实体内部标注，出现在实体首部显然是不合适的，若对该序列进行打分，I前面接O的情况要得分低才正确，这就是CRF对应的一条特征函数。如图中标注所示，CRF具有两个特征函数，其中是转移特征函数，依赖于当前和前一个位置，表示从上一个位置的标记转移到当前位置标记的概率，如上一个标签是“O”，那下一个标签是“I”的概率应当为0。是状态特征函数，依赖于当前位置，表示当前位置出现标记的概率，如在位置出现“B”标签的概率为1/3。

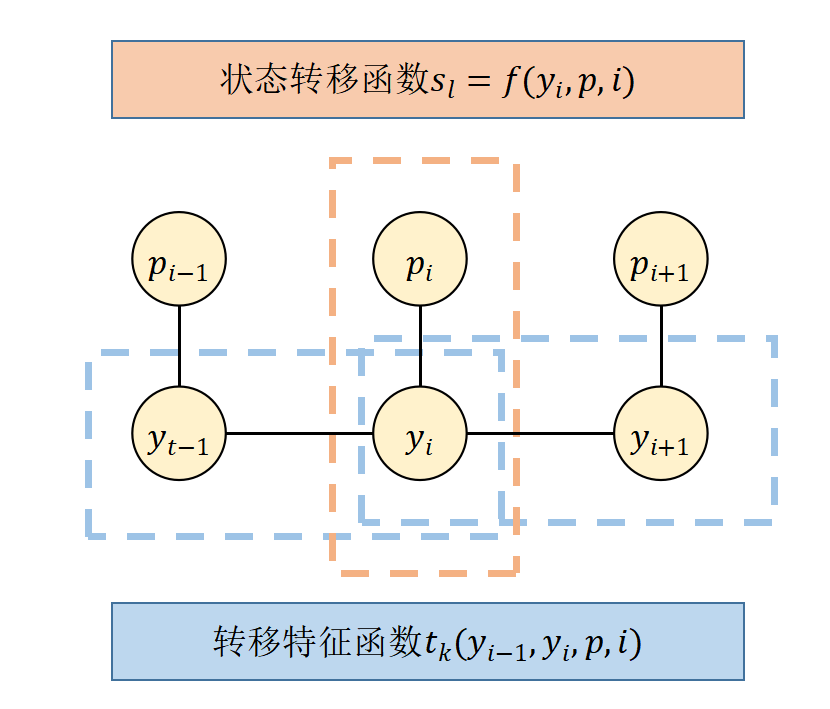


图3.3 线性链条件随机场结构示意图

在对癫痫文本进行命名实体识别时，通过双向长短期记忆网络提取了文本的特征，并计算输出每个单词对应于每个标签的概率，得到一个矩阵P。如果对每个单词都采用直接输出最大概率标签的方式，可能就会出现上例中提到的不合理的序列，如“I”标签在“O”标签的后面。所以采用CRF对文本特征提取层的输出进行约束，避免出现不合逻辑的标记序列，提高命名实体识别的准确率。将P作为CRF层的输入，转移矩阵T由CRF自动学习获得。整个输出序列Y的得分公式和概率公式如下：

（3.4）

（3.5）

其中表示标签从转移到的概率，表示第i个词语标记为的概率。为所有可能序列的集合，为正确的标签序列。损失函数如下：

（3.6）

最终可能输出的序列为预测得分最高的：

（3.7）

# 4 实验过程与结果分析

## 4.1数据来源

本文有关癫痫文本的数据集EPILEPSY主要来源于已有的资料信息，从文献和指南等爬取文本后经过预处理所得。数据集具体划分如下表：

表4.1 数据集划分

|  |  |
| --- | --- |
|  | 文本数据 |
| 训练集 | 3500条 |
| 验证集 | 1714条 |
| 测试集 | 1357条 |

EPILEPSY的文本标注主要使用BIO标记法。BIO标记规范为：B表示一个实体的头部，I是实体中间，O是非实体标签。根据癫痫医学文本中蕴含的并发症名、药物名、癫痫分类名、病因、诊疗方法等实体，华西专家提供了实体标签和标注字典，本文以BIO为标注规范通过字典对数据集进行实体标注。相关实体定义及标注如下：

表4.2 癫痫实体定义及对应标注表

|  |  |  |  |
| --- | --- | --- | --- |
| 实体名称 | 解释 | 实体头部标记 | 实体内部标记 |
| comorbiditity | 并发症名 | B-comorbiditity | I-comorbiditity |
| medication | 药物名 | B-medication | I-medication |
| classification | 癫痫分类名 | B-classification | I-classification |
| criteria | 标准名 | B-criteria | I-criteria |
| diet | 饮食 | B-diet | I-diet |
| differentia | 诊断 | B- differentia | I- differentia |
| epilepsy | 癫痫病 | B-epilepsy | I-epilepsy |
| etiology | 病因名 | B-etiology | I-etiology |
| immunotherapy | 免疫疗法 | B-immunotherapy | I-immunotherapy |
| method | 医学疗法 | B-method | I-method |
| psychotherapy | 心理疗法 | B-psychotherapy | I-psychotherapy |
| other | 其他治疗方法 | B-other | I-other |
| prognosis | 预后（医学术语） | B-prognosis | I-prognosis |
| severity | 严重程度 | B-severity | I-severity |
| surgery | 外科学术语名 | B-surgery | I-surgery |
| tool | 工具 | B-tool | I-tool |
| women | 妊娠相关术语名 | B-women | I-women |

通过字典和标注规范对数据集进行人工标注和检查。人工标注结果经过了华西医学专家的交叉验证，保证了数据的科学合理性。

## 4.2评估标准

要验证模型的有效性就需要对其进行全面评估，现在常用的评估指标有Accuracy、Precision、Recall和F1得分值。其中Accuracy是指准确率，在数据不均衡的情况下，往往具有欺骗性。例如本文实验是针对癫痫文本做命名实体识别，更关注的是实体标签即正样本，当数据集中正样本个数较少，负样本个数较多时，准确率往往也很高但这并不能说明模型实体识别效果好。因此本次实验中选取的评价指标为Precision、Recall和F1得分值，相关定义如下：

分类模型可能出现的结果如下表：

表4.3 预测情况表

|  |  |  |
| --- | --- | --- |
| 实际情况 | 预测情况 | |
| 正样本 | 负样本 |
| 正样本 | TP | FN |
| 负样本 | FP | TN |

Precision也称作精确率，表示模型预测为正的样本中有多少是真正的样本：

（4.1）

Recall指召回率，也称作查全率，表示模型覆盖到了多少正样本：

（4.2）

F1得分值是综合了Precision和Recall的一个指标：

（4.3）

## 4.3实验设置

论文实验相关环境配置如表4.4所示：

表4.4 环境配置表

|  |  |
| --- | --- |
| 环境 | 配置 |
| 系统环境 | Ubuntu 20.04.4 LTS |
| GPU型号 | RTX 2080 Ti |
| Anaconda版本 | Anaconda 4.10.3 |
| Tensorflow-gpu | 1.11.0 |
| Python版本 | Python 3.6 |

模型的主要参数设置如表4.5所示，内部参数通过随机初始化获得，优化器选用Adam算法：

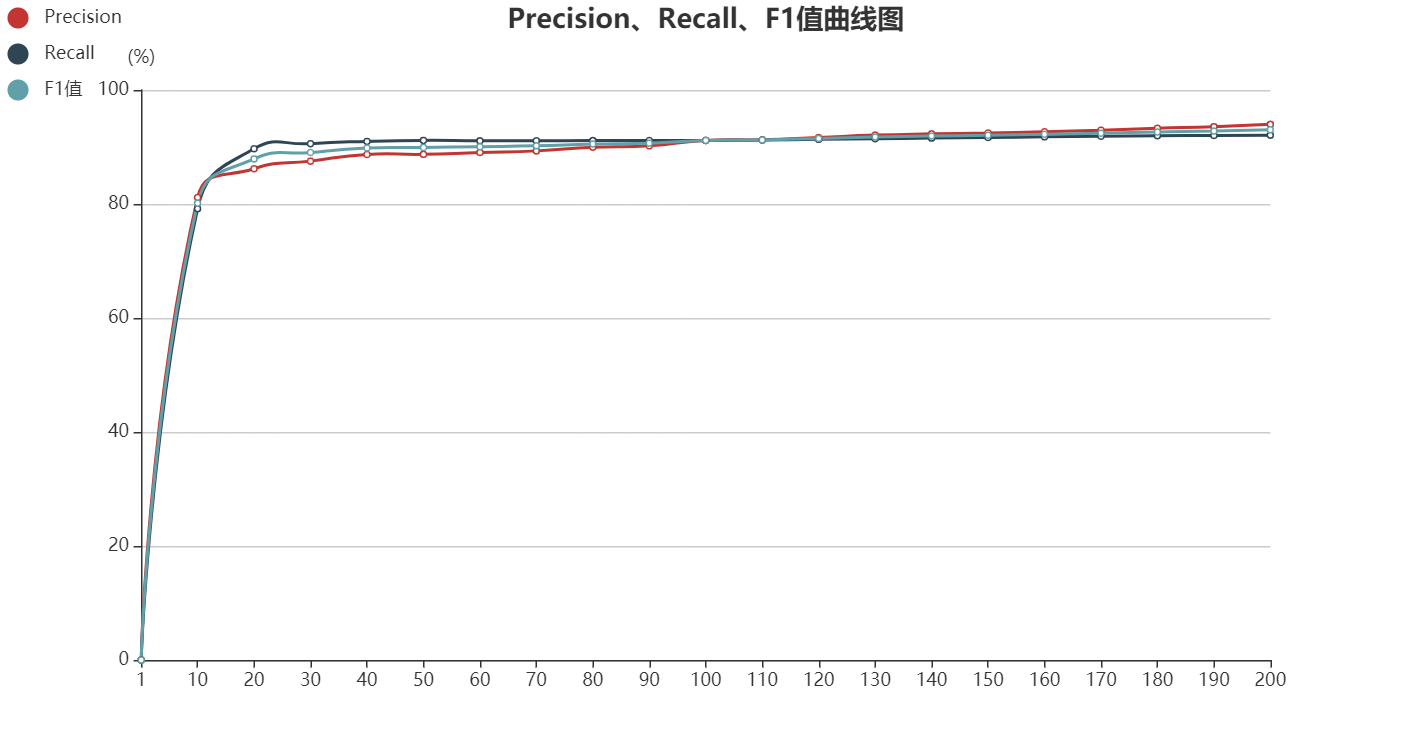
表4.5 模型参数表

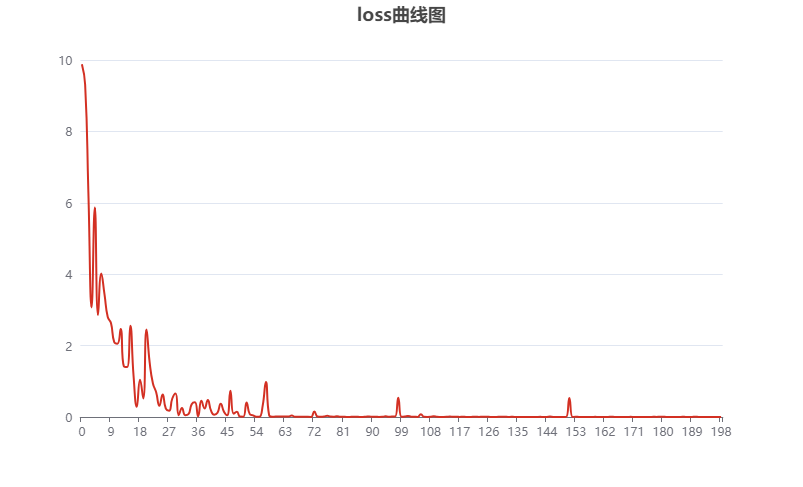
|  |  |  |
| --- | --- | --- |
| 参数名 | 参数解释 | 参数取值 |
| token\_embedding\_size | BIOBERT生成的词向量维度 | 768 |
| pos\_embedding\_size | 词性向量维度 | 30 |
| chunking\_embedding\_size | 组块标注向量维度 | 10 |
| max\_seq\_length | 最大句子长度 | 256 |
| batch\_size | 批处理大小 | 16 |
| learning\_rate | 学习率 | 1e-5 |
| num\_train\_epochs | 迭代次数 | 200 |
| dropout\_rate | 避免过拟合 | 0.5 |
| clip | 梯度裁剪 | 0.5 |
| warmup\_proportion | 慢热学习率 | 0.1 |
| lstm\_size | BiLSTM隐藏层维度 | 128 |
| num\_layers | 隐藏层层数 | 1 |

## 4.4 实验对比及结果分析

4.4.1 EPIMT-NER模型实验结果及分析

由图4.1和图4.2可以看出，随着模型迭代次数的增加，模型的loss不断下降，精确率、召回率和F1值不断增加。在10个epoch之后，模型的precision、recall和F1值均已经超过了80%，200个epoch后趋于收敛。最终模型在癫痫医学文本数据集EPILEPSY上的准确率为93.98%，召回率为92.08%，F1值为93.02%。

图4.1 Precision、Recall、F1值曲线图

图4.2 loss曲线图

与模型的总体识别结果相比，实体级别的结果能更好地评价模型。以下列出部分实体级别的结果，包含各实体的精确率、召回率和F1值。

表4.6 部分实体的识别结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | precision | recall | F1 |
| classification | 97.14% | 98.08% | 97.61% |
| comorbiditity | 89.77% | 95.18% | 92.40% |
| epilepsy | 97.24% | 97.78% | 97.51% |
| etiology | 78.95% | 75.00% | 76.92% |
| medication | 91.11% | 75.93% | 82.83% |
| method | 96.30% | 86.67% | 91.23% |
| surgery | 100.00% | 100.00% | 100.00% |
| women | 93.65% | 93.65% | 93.65% |

从表4.6中可以看出，部分实体如classification、epilepsy、method、surgery等实体识别效果非常好，三项指标都达到了95%以上，surgery实体类甚至达到了100%。这说明这几类实体的数据量充足，模型得到了非常好的学习。同时，这几类实体中数据的特征比较明显，边界确定效果较好，最终命名实体识别的效果也很好。而少数实体如etiology三项指标都不算很高的原因是数据集中相关实体数目较少，模型没有更好的学习到实体特征，可以考虑在后续工作中增加数量集，提高对该类实体的识别效果。

4.4.2 特征引入对模型性能的影响

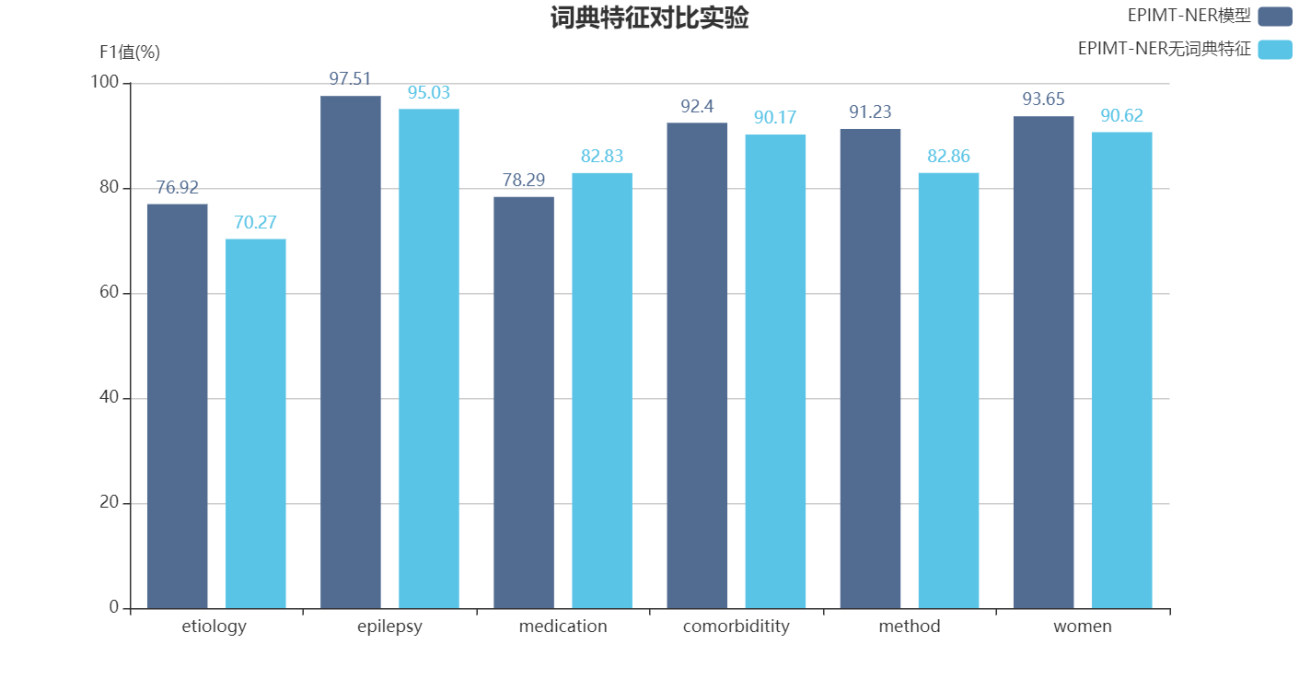
为了验证增加特征向量对解决目标问题的意义，本文将以BiLSTM-CRF模型、BioBERT-BiLSTM-CRF模型为基准分别进行一组对比实验。通过分析增加词性、组块和词典特征前后评估指标的变化来分析该模块对模型的影响。对比实验结果如表4.7所示。

表4.7 增加特征向量前后的对比实验结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| BiLSTM-CRF | 88.69% | 76.07% | 81.89% |
| feature+BiLSTM-CRF | 90.89% | 77.31% | 83.55% |
| BioBERT-BiLSTM-CRF | 90.15% | 90.98% | 90.56% |
| EPIMT-NER | 93.98% | 92.08% | 93.02% |

通过表4.7实验结果对比可以看出，在BiLSTM-CRF与feature+BiLSTM-CRF和BioBERT-BiLSTM-CRF与EPIMT-NER模型这两组对比实验中，加入词性标注和组块分析的结果相较之前有了明显的提升。第一组的实验中平均精确率提高了2.2%、召回率提高了1.24%，F值提高了1.66%。第二组的实验中平均精确率提高了3.83%、召回率提高了1.1%，F值提高了2.46%。由此可以分析出，增加词性与组块分析特征后，模型覆盖正确实体的能力明显变强，召回率方面有了极大的提升，模型的精确度和F1值也有了不错的提高。实验结果表明该模块能有效增强模型命名实体识别能力，提高识别效果。

通过对癫痫医学文本数据的特点进行分析，EPIMT-NER模型引入了词典特征以提高对特定实体的识别效果。为了验证词典特征对特定实体识别的有效性，本文增加EPIMT-NER和EPIMT-NER无词典特征的对比实验，主要比较对比实验中特定实体的F1值的变化。图4.3展示了部分特定实体的实验结果。

图4.3词典特征对比实验结果

从图4.3中对比实验的结果可以看出，针对癫痫医学文本的特定实体，增加了词典特征后，F1值都有了明显的提升，证明了词典特征能有效提高癫痫医学文本的命名实体识别。

4.4.3 词向量对模型性能的影响

为了验证前文所提的BioBERT获取癫痫医学文本词向量的有效性，本文分别使用Word2Vec、BERT、BioBERT生成词向量，再送入文本特征层和CRF层进行训练，将三个实验作为对比实验进行结果分析。本组对比实验都采用增加了特征向量模块的实验结果。

表4.8 验证BioBERT的有效性的对比实验结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| BiLSTM-CRF | 91.33% | 79.56% | 85.04% |
| BERT-BiLSTM-CRF | 93.07% | 91.53% | 92.29% |
| EPIMT-NER | 93.98% | 92.08% | 93.02% |

通过Word2Vec、BERT、BioBERT获取词向量的三个对比实验模型结果如表4.8所示，BiLSTM-CRF在精确率、召回率和F1值三个评价指标中都取得了最低分，其中精确率为91.33%，召回率为79.56%，F1值为85.04%。通过对实验结果进行统计分析，该对比实验中召回率低的原因有：数据集中个别标签的特征样本数量较少，特征学习的效果较差，导致个别标签的识别结果不好；BIO标注实体的边界问题导致实体识别数目较少。这也可以验证上文所述Word2Vec生成词向量时无法解决一词多义问题，产生的词向量不如后面两个预训练模型丰富。BERT模型具有双向语义表征，能获取到更丰富的动态词向量，具有很强大的拟合能力，因此可以用较少的标注数据获得好的识别效果。从结果可以看出，平均精确率为93.07%，召回率为91.53%，F1值为92.29%，精确率召回率都有了明显的提升，在标签特征样本较少的实体标签中也取得了不错的结果，更能说明BERT的强大。而通过BERT-BiLSTM-CRF与EPIMT-NER模型的实验结果对比可以看出，EPIMT-NER模型的平均精确率为93.98%，召回率为92.08%，F1值为93.02%。相较于BERT在平均精确率方面提升了0.91%，召回率提升了0.55%，F1值也高了0.73%，说明采用生物语料库预训练的BioBERT对癫痫医学文本的识别能力更强，覆盖正确实体的能力更好。综合F1值结果来看，针对本文数据集EPILEPSY而言，EPIMT-NER模型更适合做癫痫医学文本的命名实体识别，也通过实验数据验证了这一理论的有效性。

4.4.4 探究在公开数据集上模型的性能

EPIMT-NER模型主要结合癫痫医学文本数据特点设计，并在癫痫数据集EPILEPSY上完成实验和相关对比实验，均取得了不错的效果。为了探究模型在其他领域数据集中的效果，本文在公开数据集WikiANN中进行探究实验。WikiANN数据集是基于维基百科构建的，主要定义了三类实体：人名、地名和组织名。标签采用BIO标注格式。数据集分布如表4.9所示：

表4.9 WikiANN数据集分布

|  |  |
| --- | --- |
|  | 实体数目 |
| 训练集 | 180393 |
| 验证集 | 90535 |
| 测试集 | 90325 |

针对WikiANN公开数据集，设置EPIMT-NER和BERT-BiLSTM-CRF+feature的对比实验。实验结果如表4.10所示：

表4.10 EPIMT-NER和BERT-BiLSTM-CRF+feature的对比实验结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| EPIMT-NER | 84.69% | 87.86% | 86.24% |
| BERT-BiLSTM-CRF+feature | 92.62% | 95.04% | 93.82 |

可以看出，针对WikiANN公开数据集而言，BERT-BiLSTM-CRF的效果远远好于EPIMT-NER。这是因为EPIMT-NER模型采用预训练模型BioBERT生成词向量，而BioBERT是在大量生物语料中训练的，更适用于生物医学领域。而BERT是基于维基百科进行预训练的，更适合于通用领域。

为了探究EPIMT-NER模型中引入的特征向量模块对公开数据集WikiANN的有效性，本文设置EPIMT-NER和BioBERT-BiLSTM-CRF的对比实验，实验结果如表4.11所示：

表4.11 EPIMT-NER和BioBERT-BiLSTM-CRF的对比实验结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| EPIMT-NER | 84.69% | 87.86% | 86.24% |
| BioBERT-BiLSTM-CRF | 83.76% | 87.27% | 85.48% |

从表4.11中可以看出增加的特征向量模块对WikiANN数据集的命名实体识别有一定的作用，但是效果提升不如在EPILEPSY数据集中进行的实验明显。

结合以上对比实验综合分析可以看出，EPIMT-NER模型可以有效解决癫痫医学文本命名实体识别标注的问题，相较现有的其他方法有了明显的提升效果。同时，EPIMT-NER模型具有一定的迁移能力，但主要还是针对于癫痫医学文本数据集，在其他数据集上效果提升不明显，甚至低于其他更适合的模型。

# 5 总结和展望

## 5.1 总结

癫痫作为神经科的第二大疾病，致死致残率极高，还会严重影响女性患者的生育能力和后代健康。通过对癫痫领域文本的信息进行抽取有利于构建一个学习型的知识图谱，并在此基础上开发出智慧型诊疗系统来辅助临床决策、造福女性患者。命名实体识别是信息抽取的子任务，本论文重点对癫痫文本的命名实体识别方法进行了研究与分析，主要工作如下：

* 提出并构建了EPIMT-NER模型，利用BioBERT获取动态词向量，解决了一词多义的问题，能获取上下文信息，提高模型的精度。同时，它采用生物医学领域的语料进行训练，更适用于本次癫痫医学文本的实体识别。将词向量输入到BiLSTM层进行文本特征提取，最后通过CRF层对提取结果进行优化输出。
* 增加词性、组块和词典特征模块，获取词性标注、组块分析和词典特征的标注向量，将其与词向量进行拼接共同输入到BiLSTM层进行特征提取，加强了对癫痫医学文本信息的针对性，提高了模型的识别效果。
* 设置了对比实验，验证了EPIMT-NER模型的有效性。根据EPIMT-NER模型的实验结果可知，EPIMT-NER模型的平均精确率为93.98%，平均召回率为92.08%，平均F1值为93.02%。在分别用Word2Vec、BERT和BioBERT获取词向量的对比实验中，EPIMT-NER模型较Word2Vec对比实验来说三项指标都有了明显的提升，精确率提升了2.65%，召回率提升了12.52%，F1值提升了7.98%；较BERT模型生成词向量的对比实验，EPIMT-NER模型在召回率方面有了明显的提升，验证了预训练模型选择BioBERT的有效性。在BiLSTM-CRF与POS+Chunk ing+BiLSTM-CRF的实验中后者精确率提高了0.49%、召回率提高了0.82%，F值提高了0.66%。BioBRET-BiLSTM-CRF与EPIMT-NER模型的实验中EPIMT-NER模型的精确率提高了3.83%、召回率提高了1.1%，F值提高了2.46%。最终通过实验结果综合分析证明本文提出的EPIMT-NER模型有效提升了对癫痫医学文本数据集进行命名实体识别的效果。

在实验过程中遇到了很多问题，以下对其中一些问题和解决方案进行记录说明：

* BERT的分词产生的问题。BERT分词有两个过程，首先是按空格粗分，然后对token进行细粒度划分，在细粒度划分的过程中会产生一些多余信息，这些信息需要我们自己去处理，在最后标记的时候去掉多余标签X，否则在命名实体识别的过程中就会产生错误，得到非常低的识别效果。
* 在加入词性向量、组块分析向量和词典特征向量时产生的问题。最初是想利用BioBERT模型生成词性向量和组块向量，结果提示资源耗尽，显存爆掉了。后来经过思考和分析，同时调用多个BioBERT模型获取词向量实在是太大了，非常不现实，后又改为用Word2Vec进行词性特征、组块分析特征和词典特征的向量获取。这几个向量并不会出现一词多义的现象，用Word2Vec直接获取也不会太影响模型的性能和效果，目前认为是最合适的办法。

总的来说，本次论文工作的开展还算顺利，完成了模型的设计构建与实现，最终也取得了不错的实验结果。我在整个过程收获了很多，同时也为进一步学习自然语言处理打下了坚实的基础。

## 

## 5.2 展望

本文提出的EPIMT-NER模型在癫痫医学文本进行命名实体识别任务中整体达到了预期效果，但仍有一些可以改进的地方：

* 模型训练的时间较长。BiLSTM的参数复杂度较高，可以考虑使用更简化的神经网络模型来减少模型训练的时间。
* 数据量相对较少。由于数据集来源于人工爬取和标注，可在之后的工作中对数据集进行不断更新，进一步提高模型识别的效果。

此外，在本次实验的基础上，未来可以进一步学习开展癫痫知识图谱构建工作。同时为知识图谱在智能医疗方面的应用打下基础。

**参考文献**

**声** **明**

本人声明所呈交本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得四川大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

本学位论文成果是本人在四川大学读书期间在导师指导下取得的，论文成果归四川大学所有，特此声明。

学位论文作者（签名）

论文指导教师（签名）

年 月 日

**致 谢**

难以想象，四年的大学生活就要接近尾声了，感觉2018年刚开学的日子也就在不久前。这四年来，我时常庆幸自己来到了四川大学，来到这个充满包容与爱的大学。它真的做到了“海纳百川，有容乃大”，也让每一个川大人将校训刻在了心里。

很感谢段磊老师的关心与指导，让我不断进步与成长；感谢我的辅导员叶庆老师，让我懂得了很多人生的道理，走过了迷茫与低沉的时期；感谢孙毓同学的陪伴和鼓励，让我的大学生活更加丰富有趣；感谢室友一直以来的包容和陪伴，很幸运遇到你们。还有年级班委的小伙伴、馨心社的小可爱以及没有提到的我的朋友们，真的很感谢和你们相处的时光，我们一起学习，一起玩耍，一起为同一个目标而努力。虽然那些日子再也回不来了，但它深深地留在了我的记忆里，留在了江安校区的每一处风景里。

最后，感谢一直努力没有放弃的自己，感谢那些早起的清晨，感谢那些为学习熬过的夜。

希望母校越来越好，希望四川大学计算机学院越来越好，希望四川大学的每一位老师和同学都心想事成，万事如意！

1. [] Rau L F . Extracting company names from text[C]//Proceedings of the 7th IEEE Conference on Artificial Intelligence Applications. Los Alamitos: IEEE, 1991: 29-32. [↑](#endnote-ref-0)
2. [] Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History[C]//Proceedings of the 16th International Conference on Computational Linguistics, 1996: 466-471. [↑](#endnote-ref-1)
3. [] 江千军, 桂前进, 王磊, 徐瑞翔, 王京景, 麦立, 许水清. 命名实体识别技术研究进展综述[J]. 电力信息与通信技术, 2022, 20(02): 15-24. [↑](#endnote-ref-2)
4. [] Todorovic B T, Rancic S R, Markovic I M, et al. Named entity recognition and classification using context hidden Markov model[C]//Proceedings of 2008 9th Symposium on Neural Network Applications in Electrical Engineering, 2008: 43-46. [↑](#endnote-ref-3)
5. [] Adam L. Berger, Stephen Della Pietra, Vincent J. Della Pietra. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1): 39-71. [↑](#endnote-ref-4)
6. [] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning, 2001: 282-289. [↑](#endnote-ref-5)
7. [] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science, 2015. [↑](#endnote-ref-6)
8. [] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing, 2014: 1532-1543. [↑](#endnote-ref-7)
9. [] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018. [↑](#endnote-ref-8)
10. [] Wang Zi-niu, Jiang Meng, Gao Jian-ling, et al. Chinese named entity recognition method based on BERT[J]. Computer Science, 2019, 46(S2): 138-142. (in Chinese) [↑](#endnote-ref-9)
11. [] 许力, 李建华. 基于BERT和BiLSTM-CRF的生物医学命名实体识别[J]. 计算机工程与科学, 2021, 43(10): 1873-1879. [↑](#endnote-ref-10)
12. [] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013. [↑](#endnote-ref-11)
13. [] Jinhyuk Lee et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2019. [↑](#endnote-ref-12)