东 北 大 学

**研 究 生 课 程 作 业**

评分

课程科目： 语言分析与机器翻译

课程编号：

阅 卷 人：

学生学院： 计算机科学与工程学院

学生专业： 计算机科学与技术

学生姓名： 王昊天

学生学号： 20182853

东北大学研究生院

2021年11月30日

基于BiLSTM CRF对人民日报1998数据集的命名体识别

1. 内容

1、实现LSTM CRF模型

2、处理人民日报1998数据集至一定格式

3、对模型进行训练

4、对模型进行测试

1. 模型
2. LSTM

循环神经网络（RNN）是一种可以感知历史信息的循环神经网络，RNN能够捕捉非线性短期时间依赖关系。可以区分全连接和部分连接的RNN。第一个RNN是由Williams和Zipser在1980 年代后期开发的，当时神经网络结构的发展带来了该领域的众多基础性贡献。目前的研究中，应用 RNN 在语音识别，语言建模，翻译，图片描述等问题上已经取得一定成功，RNN 的关键点之一就是他们可以用来连接先前的信息到当前的任务上，例如使用过去的视频段来推测对当前段的理解。但随着时间的推移，RNN最终会导致梯度消失或爆炸。

LSTM由Hochreiter和Schmidhuber于1997年提出，旨在解决RNN模型中的“梯度消失”和“梯度爆炸”问题。LSTM模型由一个或多个内存模块作为其基本单元组成。每个模块都包含存储单元和门来控制系统的信息流。因此，LSTM近年来在各个领域得到了广泛的应用，并取得相当巨大的成功。

在此，LSTM算法的输入向量表示为X=(x1,x2,…,xn),xi∈RT,i=1,2,…,n,n表示输入向量的维数，T表示时间序列的时间滞后性，Y=(y1,y2,…,yn)表示输出序列，LSTM学习过程描述为：

LSTM 单元中包含遗忘门、输入门和输出门，遗忘门对前一单元信息进行选择性遗忘， zf以前一单元信息和当前状态作为输入，由 sigmoid 函数输出范围 0~1 值，此值为保留传递信息的百分比。当前单元输入信息比例由输入门控制，如式(2)所示，zi 为保留信息所占的比重，将保留信息和新信息进行加权处理作为当前单元的状态；输出门决定有多少信息被输出，将当前单元的部分信息传递给后边的单元。

1. CRF

NER是一个分类问题，但因为一个词在不同上下文中含义是不同的，所以NER是一个基于上下文的分类问题，一个序列问题。传统方法中解决序列问题比较成功的方法有两个HMM和CRF。相对于HMM，CRF有两个优势CRF不仅可以利用相邻词之间的关联，还可以设置更大的邻域CRF可以方便融合多特征

CRF，英文全称为conditional random field, 中文名为条件随机场，是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型，其特点是假设输出随机变量构成马尔可夫（Markov）随机场。









v较为简单的条件随机场是定义在线性链上的条件随机场，称为线性链条件随机场（linear chain conditional random field）. 线性链条件随机场可以用于序列标注等问题，而本文需要解决的命名实体识别(NER)任务正好可通过序列标注方法解决。这时，在条件概率模型P(Y|X)中，Y是输出变量，表示标记序列（或状态序列），X是输入变量，表示需要标注的观测序列。学习时，利用训练数据 集通过极大似然估计或正则化的极大似然估计得到条件概率模型p(Y|X)；预测时，对于给定的输入序列x，求出条件概率p(y|x)最大的输出序列y0.

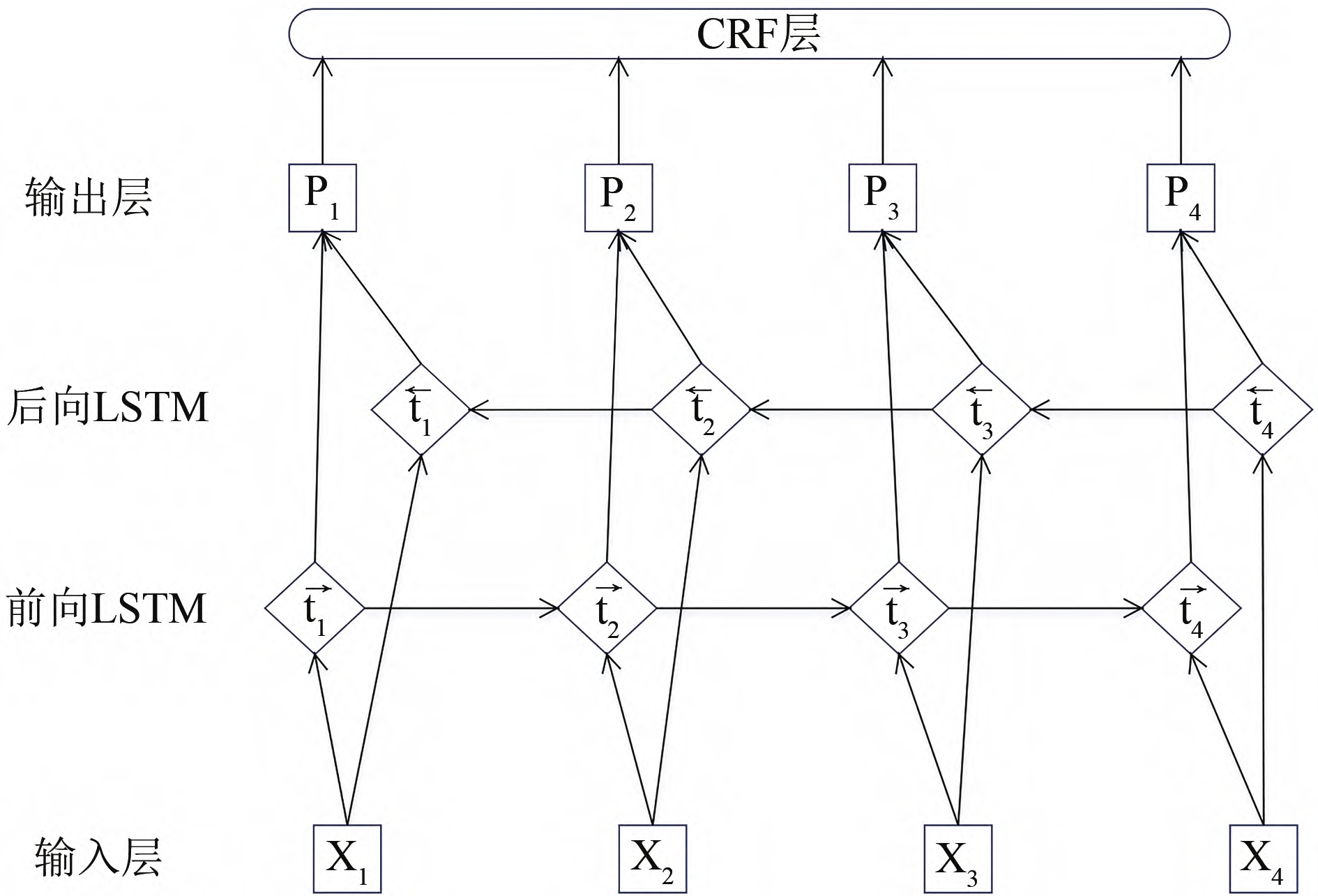
1. BiLSTM-CRF

近年来，多种深度学习方法被广泛应用于命名实体识别任务中，通常采用RNN模型及其变体，RNN在理论上能够捕获远程上下文关系，但在实践中由于梯度消失或梯度爆炸而失效。因此LSTM在实际应用中经常被用到，在命名实体识别任务中，同时访问当前时刻的上下文有助于预测当前的时刻，但是LSTM的隐藏状态h（t）只接受过去的信息。因此笔者使用双向LSTM模型来给出每个状态的上下文，从左到右和从右到左使用两个独立的隐藏状态，同时捕获过去和未来的信息［18］。在模型输出层上，笔者会将捕获的上下文信息经向量化后作为标注特征，但若仅通过该模型进行实体识别则会面临标注间约束关系无法表示的问题。而用CRF模型实现中文分词，则可在具有良好学习性能的同时从一定程度上实现对生词的识别。

BILSTM-CRF模型主要包括两部分，即BILSTM层和CRF损失层。对于一个输入句子x，首先经过embedding层将每个词汇或者字符映射为一个词向量或者字符向量，然后传入BILSTM层，获得句子的前向和后向向量，接着将前向和后向向量进行拼接作为当前词汇或字符的隐藏状态向量。结果由CRF输出，损失函数的形式主要由CRF给出 在BiLSTM-CRF中，给定输入序列X，网络输出对应的标注序列y。







第一层是输入层（字嵌入层），实现文本数据向计算机可处理的向量矩阵的转换。本层通过Word2vec工具，基于训练好的模型将基于字的直接向量表示的高维稀疏向量转变为可计算词间关系的低维稠密向量。本文使用中医医案的字符序列作为输入得到字向量。

第二层是隐藏层（BiLSTM 层），实现对文本语句特征的提取。本层由两个LSTM层组成，由于对应处理向量序列次序的差异将两者区分为正序的前向层和逆序的后向层，前向隐层负责提取中医文献中每一个字的表征上文，后向隐层负责反向特征提取中医文献中每一个字的表征下文。在具体输入过程中，依次把输入向量序列的顺序和逆序序列分别作为前向和后向的输入数据。

第三层是输出层，在BiLSTM网络的输出层中为每一个输入的数据打一个标签的预测分值，接着在输出层的后面添加CRF层（逻辑回归层），它的主要作用是对语句进行序列标注，以此来增加文本之间的信息相关性。

1. 训练

以 1998 年上半年人民日报语料作为实验数据，其中后 5 个月语料作为训练数据，按惯例一月份语料为测试数据。

1. 数据预处理

PFR语料库是对人民日报1998年上半年的纯文本语料进行了词语切分和词性标注制作而成的，严格按照人民日报的日期、版序、文章顺序编排的。文章中的每个词语都带有词性标记。目前的标记集里有26个基本词类标记（名词n、时间词t、处所词s、方位词f、数词m、量词q、区别词b、代词r、动词v、形容词a、状态词z、副词d、介词p、连词c、助词u、语气词y、叹词e、拟声词o、成语i、习惯用语l、简称j、前接成分h、后接成分k、语素g、非语素字x、标点符号w）

人民日报1998数据集为分词词性的标注，无法直接用在命名体识别中，为此我从中选择了几个词性作为标注分别为："n","v","u","d","a"，其他词性均替换为“O”。然后利用分词信息，对每个字进行标注，根据字在词中的位置可分为B（开始）、I（中间）、E（结束）、S（单个）

例如：树——S，中国——BE，中华人民——BIIE。

接着与词性标注拼接，作为每个词的tag。

1. 模型训练

将语料库分组训练，每个batch包括100个句子，embedding层嵌入100维向量，BiLSTM隐藏层设置维100，训练时，出现了标签便宜，经过排查这是因为设计的损失函数除了一个len(tag)为了缩小loss，然而这样是不合适的，由于标签本来就偏向预测“O”，之后对比他人实验，决定通过减少标签种类来减缓数据不均衡。同时因为需要计算所有路径，那么在计算过程中，计算每一步路径得分之和和直接计算全局得分是等价的，就可以大大减少计算时间。

1. 结果

以2~6月5 个月语料为训练数据，1月份语料为测试数据，进行了两组实验。第1组实验是分词词性标注一体化系统，第2组采用的是传统词法处理模式，分词和词性标注分两个阶段进行。

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 精准率 | 召回率 | F1 |
| N | 93.77 | 94.41 | 94.61 |
| V | 93.05 | 94.58 | 93.81 |
| U | 97.31 | 97.71 | 97.52 |
| D | 91.82 | 94.67 | 95.22 |
| a | 91.01 | 93.48 | 92.22 |

表中列出了5种出现次数最多的词法成分相应的词性标注精度。其中歧义性较小的助词，性能最为优异；而副词和形容词的处理性能相对较低，说明这两种词性在分词过程中遇到的歧义现象较多，所以副词和形容词的歧义处理，是今后继续提升中文词法歧义处理的重点。

1. 感想

通过本次实验，更加熟悉了pytorh架构，以及lstm的原理，对于原理的掌握，自以为已经非常了解其原理，但是再实现时确是捉襟见肘，各种地方出问题，对于数据集处理，是一个大头，对数据格式的不了解，看不懂德文的结构，对于英德翻译无从下手，而人民日报98数据集也是如此，对数据集不了解，不知道数据中还有多种歧义标签，导致在数据处理以及分词是出现问题。需要对数据进行清洗才能够使用。因为对LSTM的实现不太熟悉，导致自己很难弄懂其输入输出，才意识到自己对LSTM只是似懂非懂。实践是检验真理的唯一标准，不能眼高手低，多实践才能掌握真知识。对于超参数的设置也体会到了非常麻烦，由于经验的不足，不知道超参数该设置多大的规模，loss的大小来决定优化器的学习效率。通过本次实验很有效的提高了自己的编程能力。