

**自然语言处理实验**

**(2020级计科专业)**

**计算机科学与技术学院**

**自然语言处理课程组**

**2023年6月**

目 录

[实验-1 中文分词 1](#_Toc89192771)

[1.1 实验目的 1](#_Toc89192772)

[1.2 预备知识 1](#_Toc89192773)

[1.3 实验任务 2](#_Toc89192774)

[实验-2 基于BiLSTM-CRF的命名实体识别 4](#_Toc89192775)

[2.1 实验目的 4](#_Toc89192776)

[2.2 预备知识 4](#_Toc89192777)

[2.3 实验任务 6](#_Toc89192778)

[华为云ModelArt Ascend 环境配置 8](#_Toc89192779)

[实验报告提交 8](#_Toc89192780)

# 实验-1 中文分词

## 1.1 实验目的

通过实验了解中文分词的大致思路。在实验中实现中文分词处理，可考虑使用课堂讲解过的算法（比如基于统计、基于词典的分词方法等），或者课外学习算法（比如Bi-LSTM+CRF模型等）进行分词，最后对比不同算法分词效果和性能，加深对中文分词算法的理解。

## 1.2 预备知识

中文分词指将汉字序列切分成单个词语的过程。在以英文为代表的拉丁语系中，单词之间是以空格作为自然分界符的。而中文只有字、句和段能通过明显的分界符来划界，唯独词没有一个形式上的分界符，这给中文的处理带来了独有的困难，因此中文分词技术得到了广泛研究。

理论上讲，构建一套完备的分词规则便可以将所有句子正确划分，但语言规则庞大复杂并且是动态发展的，编写这样一套规则是不现实的，因此目前主流的分词方法可以大致分为：（1）基于词典匹配的分词算法；（2）基于统计学习的分词算法；（3）基于深度学习的分词算法等。

**1.2.1 基于词典匹配的分词算法**

基于词典匹配的分词算法依赖人工建立的词库（词典）进行，包括正向最大匹配法、逆向最大匹配法以及双向最大匹配法。这里给出正向最大匹配法的大致流程：

1. 从左向右取句子的前m个字作为匹配字段(m为词典中最长词的长度)
2. 查找词典进行匹配
3. 若匹配成功，则将该字段作为一个词切分出去
4. 若匹配不成功，则将该字段最后一个字去掉，剩下的字作为新匹配字段，再次进行匹配
5. 重复上述过程，直到切分所有词为止

逆向最大匹配法则从待分词句子的末端开始，也就是从右向左开始匹配扫描，每次取末端m个字作为匹配字段，匹配失败，则去掉匹配字段前面的一个字，继续匹配。双向最大匹配法将正向最大匹配法得到的分词结果和逆向最大匹配法得到的结果进行比较，选取更为合适的作为结果。这里给出一种选择方式：

1. 如果正反向分词结果词数不同，取分词数量少的那个
2. 如果分词结果词数相同：
   1. 分词结果相同，没有歧义，返回任意一个
   2. 分词结果不同，返回其中单字数量较少的那个

**1.2.2 基于统计学习的分词算法**

将中文分词视作一个**序列标注**任务，给句子中的每个词打上合适的标签就可以完成分词任务。下面介绍本实验中使用的“BIES”标注方式：

B：代表该字是一个词的开头

I：代表该字在一个词的内部

E：代表该字是一个词的结尾

S：代表该字自己就是一个词，例如：“的”

使用上述标注方式对句子进行标注，例如：

B E / S / B E / B I E / B E

小明 / 在 / 中国 / 科学院 / 工作

序列标注方法可以使用HMM等传统统计方法或者Bi-LSTM等深度学习方法，本小节介绍一种课堂外的中文分词模型，即基于双向长短神经网络+条件随机场（Bi-LSTM+CRF）的中文分词模型，算法具体原理大家可课外自行学习（课堂给出环境配置方案以及具体源码，需要大家自行熟悉）。

## 1.3 实验任务

实验任务包括基础任务与选做任务。基础任务中需实现基于词典和基于统计的中文分词算法，完成后可获得实验课程基础分。选做任务中需要对基础任务中的分词器进行优化，选做部分的分数通过分词器在测试集上的表现决定。最终提交的实验报告中应包括基础任务完成情况与选做任务中采取的优化措施。

**1.3.1 基础任务**

1. 实现基于词典的分词算法

实验一资料包下的“Dictionary\_based”文件夹中提供了基础词典和分词算法的大致框架。分词算法的核心部分需要大家完成，实验中提供了若干测试样本用以帮助大家判断算法是否正确实现。

1. 实现基于统计的分词算法

实验中给出Bi-LSTM+CRF模型的基础实现，相关代码及说明文档位于实验一资料包下的“Bi-LSTM+CRF”文件夹下。请根据给定的实验资料中README.md文件配置相应实验环境，说明：（1）提供源码PyTorch语言编写（可根据个人掌情况用其他语言编写），默认运行版本是CPU版本；（2）如希望运行NPU版本，大家可跟任课老师联系，申请华为云资源运行（需提前统计名单：姓名+学号+个人手机号码+邮箱）；

**1.3.2 选做任务**

优化基础任务中实现的分词器，可考虑的优化方案有：

1. 修改网络结构，例如引入BERT等预训练语言模型；
2. 与命名实体识别算法相互配合，减少对命名实体的错误分割；
3. 构造合适的词典集（可扩充+人工整理）；
4. 实现新词发现（登录）功能，识别测试集中的新词（未登录词）；
5. 调整、优化模型训练过程中的超参数。

完成优化后对测试文件“Bi-LSTM+CRF/data/test.txt”进行分词，分词结果保存到.txt文件中utf-8编码，词与词之间以空格分隔，每个测试样本占一行。文件“Bi-LSTM+CRF/cws\_result.txt”中给出了输出示例。提交分词结果后，依据单词级别的F1-score进行评判，决定选做部分的实验分数。

单词级别的F1-score的计算方式如下：

Gold: 共同 创造 美好 的 新 世纪 —— 二○○一年 新年 贺词

Hypothesis: 共同 创造 美 好 的 新 世纪 —— 二○○一年 新年 贺词

Precision = 9 / 11 = 0.818

Recall = 9 / 10 = 0.9

F1-score = 2\*Precision\*Recall/(Precision+Recall)=0.857

# 实验-2 基于BiLSTM-CRF的命名实体识别

## 2.1 实验目的

通过实验达到：⑴ 掌握序列标注、命名实体识别的理论基础及常规解决方案；⑵ 加深对循环神经网络、LSTM、条件随机场等模型基本原理的理解；⑶ 掌握基于TensorFlow等框架对循环神经网络、长短期记忆网络的实现方法。

## 2.2 预备知识

**2.2.1 命名实体识别**

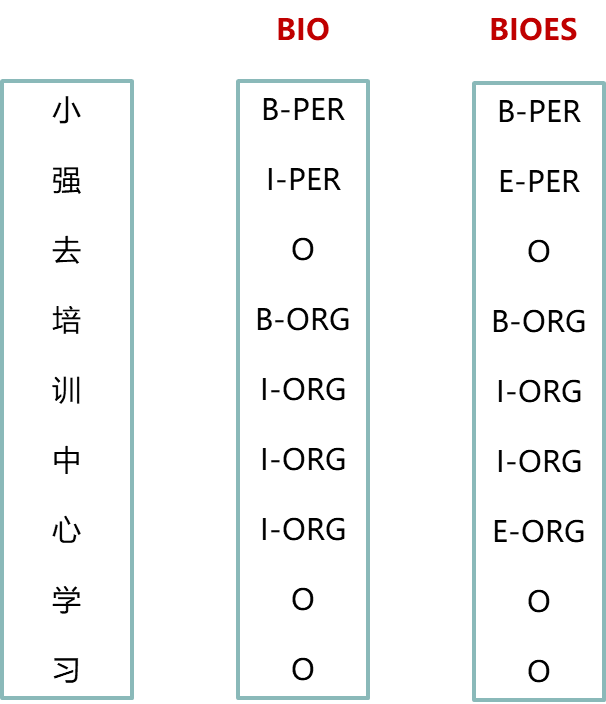
命名实体识别（Named Entity Recognition, NER），是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等，以及时间、数量、货币、比例数值等文字。

NER是一种序列标注问题，因此数据的标注方式也遵照序列标注问题。命名实体识别不仅要找出实体的位置，还要对实体进行分类。位置和类别通过标签来表达，数据标注格式主要是BIO和BIOES两种。

BIO标注法中B表示实体开始，I表示实体内部，O表示实体外部；

BIOES标注法中B表示实体开始，I表示实体内部，E表示实体结束，S表示单个词形成实体，O表示实体外部。

举例如下：

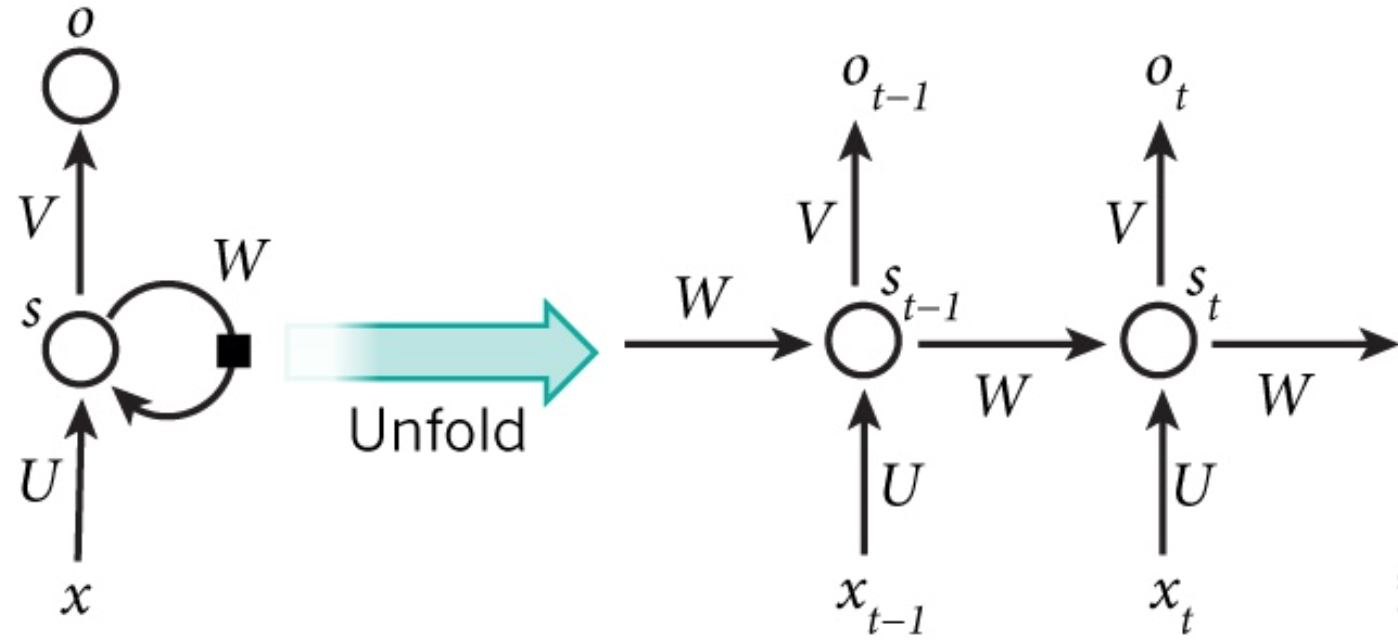


**图2.1 标注方案对比**

**2.2.2 Bi-LSTM**

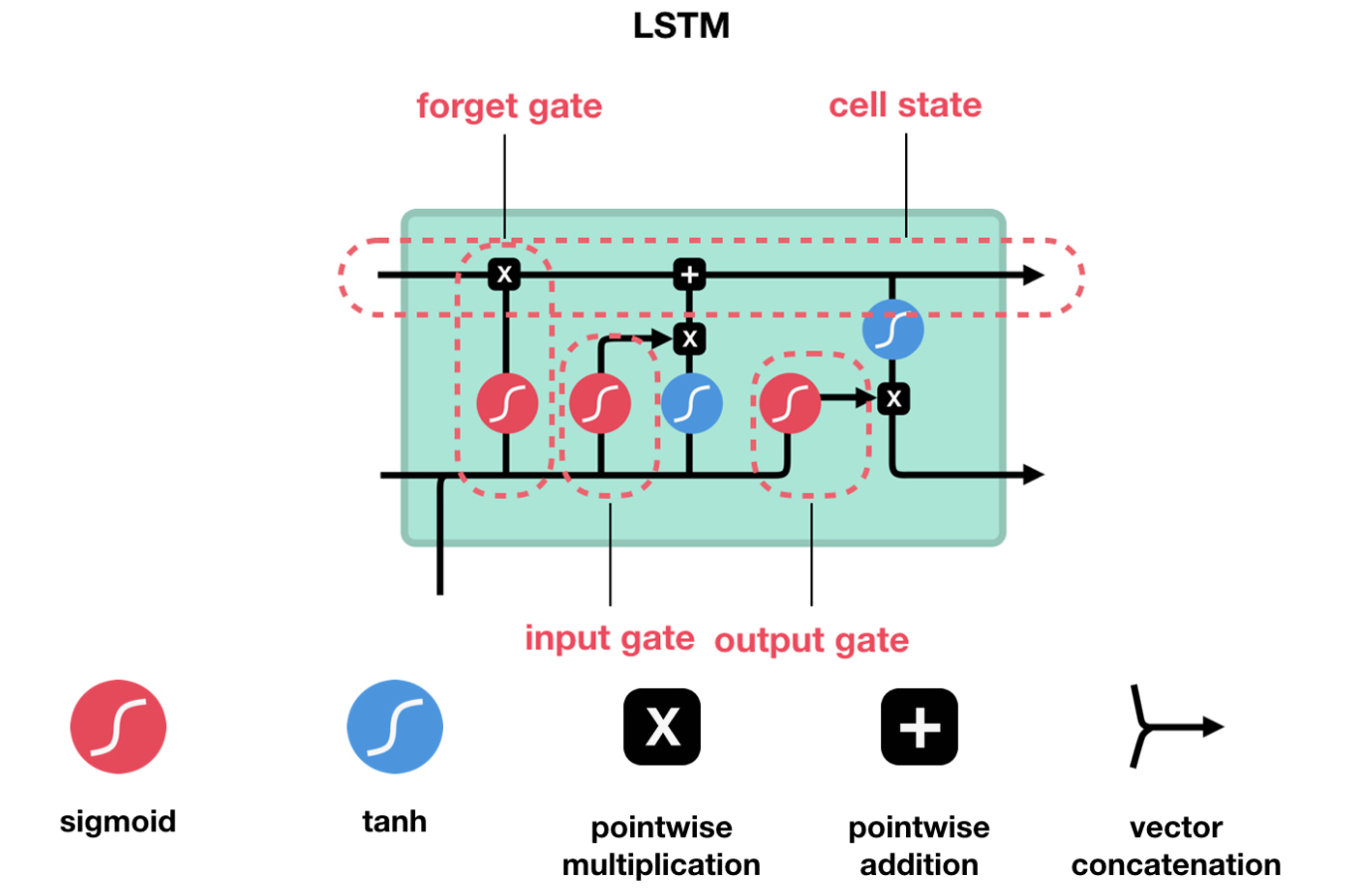
循环神经网络（RNN）被广泛运用于序列任务或时序任务，且具有良好的基线效果。在命名实体识别任务上，此类神经网络通常能够兼顾预测准确率与运行速度，得到了广泛运用。

RNN的基本结构可由下图表示：



* 是第层的输入，它可以是一个词的one-hot向量，也可以是概率分布表示；
* 是第层的隐藏状态，它负贵整个神经网络的记忆功能。由上一层的隐藏状态和本层输入共同决定，,通常是一个非线性的激活函数，比如tanh或ReLU。由于每一层的都会向后一直传递，所以理论上能够捕获到前面每一层发生的事情（但实际中太长的依赖很难训练）。
* 是第t层的输出，比如我有预测下一个词是什么时，就是一个长度为的向量，是所有词的总数，表示下一个词是的概率。最后用softmax对这些概率进行归一化
* 每一层的参数是共享的，这样极大地缩小了参数空间；每一层并不一定都得有输入和输出，比如对句子进行情感分析时只需要最后一层给一个输出即可，核心在于隐藏层的传递。

LSTM作为RNN的变种之一，通过引入门机制缓解了可能出现的梯度消失与梯度爆炸问题，其单元结构如图2.2所示。



**图2.2 LSTM单元结构**

双向LSTM即从前后两个方向传递时序信息的隐含层，LSTM的详细推导及理解可参考资料 ：

中文：<https://cloud.tencent.com/developer/article/1661253>

英文：<https://colah.github.io/posts/2015-08-Understanding-LSTMs/>

## 2.3 实验任务

实验基础任务部分要求构造一个命名实体识别（NER）模型，除了基本的预测功能外，能够对测试集进行批量预测并将测试结果保存为文件。

**2.3.1 基础任务**

1. 实现基于Bi-LSTM+CRF的命名实体识别算法

实验二资料包下的“RMRB\_NER\_CORPUS.txt”文件中提供了基于人民日报的NER标注数据，需要对数据集进行合理比例的划分，使其可以用于训练命名实体识别模型。

分词实验与命名实体识别实验所采用的模型有一定交集，因此除了自主实现模型以外，还可以参考实验1必做项中给出的Bi-LSTM+CRF标准实现并对其进行部分修改。若选择对实验1必做项中的Bi-LSTM+CRF模型进行修改，主要需要修改的部分包括**数据预处理**、**模型的输入输出层**。

1. 尝试用命名实体识别算法提升分词模型的性能

命名实体识别结果将对特定名词的识别产生提升效果，请你尝试利用NER模型结果优化实验一中的分词结果。请自行设计融合策略，并在实验报告中进行说明。

**2.3.2 选做任务**

为了进一步优化实验一的分词结果，可以从以下角度进行改进：

（1）优化命名实体识别模型，可考虑的优化方案有：

1. 修改网络结构，例如引入BERT等预训练语言模型；
2. 调整、优化模型训练过程中的超参数。

（2）数据增强

实验二提供的人民日报语料与分词所采用的语料并不一定是同分布的，你可以自行搜集更为合适的数据集进行训练。

（3）调整融合策略

# 华为云ModelArt Ascend 环境配置

华为云ModelArts是面向开发者的一站式AI开发平台，帮助用户快速创建和部署模型，管理全周期AI工作流，本次实验可选择华为云ModelArt Ascend作为实验环境，具体步骤如下：

1. 创建OBS桶，上传实验数据
   1. 登录OBS平台并创建桶；
   2. 新建“NLP”文件夹；
   3. 本地解压实验数据，上传至云端“NLP”文件夹；
2. 创建notebook开发环境
   1. 登录华为云modelarts控制台；
   2. 点击开发环境->notebook->创建，并输入notebook描述；
   3. 选择Ascend 910环境；
   4. notebook存储位置选择之前创建的“NLP”文件夹

等待notebook实例创建完成后点击“打开jupyterlab”，可以在右侧文件管理栏中看到包含实验数据的“NLP”文件夹，启动notebook即可开始实验。

# 实验报告提交

撰写本次实验报告，其内容至少包括问题描述、系统设计、系统实现和实验小结。实验报告需要按照规范格式要求规范排版。

1. 按照公告的时间及时提交电子档实验资料，所有资料存储于每位同学自己的相应文件夹下，**其文件夹名称格式为“姓名\_学号”。如：李某某\_ U201816666**
2. 资料至少包括实验报告、实验预测结果(**测试集：test\_data(需要提交分词结果).txt**)；
3. **预测结果请提交至：**[**http://211.67.20.220:55556/**](http://211.67.20.220:55556/)**（请使用校园网访问）**；预测结果文件命名规则:

“姓名\_学号-n.txt”(n=1,2)， 以两次提交最高分认定为最终成绩；

1. **实验报告命名规则为：“姓名\_学号.doc”**

最终的分词结果所达到的性能将以较大比重影响实验课得分。