

**课 程 实 验 报 告**

**课程名称： 自然语言处理实验**

**专业班级： 计算机科学与技术202207**

**学 号： U202215561**

**姓 名： 瞿明睿**

**指导教师： 魏巍**

**报告日期： 2025年 5 月 5 日**

**计算机科学与技术学院**

目 录

[1 中文分词 2](#_Toc458159879)

[1.1 问题描述 2](#_Toc458159880)

[1.2 基础模块 2](#_Toc458159882)

[1.3 系统实现 2](#_Toc458159883)

[1.4 实验小结 2](#_Toc458159884)

[2 中文命名实体识别 2](#_Toc458159885)

[2.1 问题描述 2](#_Toc458159886)

[2.2 基础模块 2](#_Toc458159887)

[2.3 系统实现 2](#_Toc458159888)

[2.4 实验小结 2](#_Toc458159889)

[参考文献 2](#_Toc458159900)

[附录A 中文分词实现的源程序 2](#_Toc458159901)

[附录B 中文命名实体识别的源程序 2](#_Toc458159902)

# 1 中文分词实现

## 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。

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。

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。

## 1.1.1 ××××××

××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××,×××××××××××××××××××××××××

××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××,×××××××××××××××××××××××××

## 1.2 基础模块

××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××× (如图1-1所示)



图1-1□××××××××××

## 1.3 系统实现

××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××××× (如表1-1所示)

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）；

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。

表1-1□××××××××××

|  |  |  |  |
| --- | --- | --- | --- |
| ××× | ××× | ××× | ××× |
| ××× | ×× | ×× | ×× |
| ××× | ×× | ×× | ×× |
| ××× | ×× | ×× | ×× |
| ××× | ×× | ×× | ×× |

.......

## 1.4 实验小结

# 2 中文命名实体识别实现

## 2.1 问题描述

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 基础模块

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。

## 2.3 系统实现

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

## 2.4 实验小结

# 参考文献

[1] 郑捷著. NLP汉语自然语言处理---原理与实践. 电子工业出版社

# 附录A 中文分词实现的源程序

import torch

import torch.nn as nn

from torchcrf import CRF

from torch.nn.utils.rnn import pack\_padded\_sequence, pad\_packed\_sequence

# 附录B 命名实体识别实现的源程序