人工智能与数据挖掘课程设计

命名实体识别研究报告

**组员信息：**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 姓名 | 学号 | 成员贡献 | 实践成绩 |
| 1 | 罗易 | 2021522009 | 数据收集、代码实现、报告撰写 |  |
| 2 | 李宁宁 | 2021522012 | 数据收集、代码实现、报告撰写 |  |
| 3 | 付天玥 | 2021522015 | 数据收集、代码实现、报告撰写 |  |

# **引言**

## 问题描述

命名实体识别（Named Entity Recognition，简称NER）是自然语言处理领域的一项重要任务，旨在识别文本中具有特定意义的实体，如人名、地名、组织名等。通过对文本进行预处理，如分词、词性标注等，NER模型可以利用深度学习算法来识别和分类文本中的实体，帮助计算机理解文本框架，为信息抽取、问答系统、语义分析等任务提供支持。

## 任务分析

首先是数据的准备，我们需要收集包含命名实体标注的语料库或文本数据集，确保数据覆盖各种类型的命名实体，然后对数据进行清洗和预处理，保证数据格式的统一。

接着进行模型的选择，根据数据规模、任务需求和计算资源选择最适合的深度学习模型，如隐马尔可夫模型（HMM）、条件随机场（CRF）、循环神经网络（RNN）、变换器（Transformer）等等。

然后搭建和训练模型，将数据集划分为训练集和验证集，对模型进行训练，调整模型超参数，如学习率、正则化项等，以优化模型性能。

最后对模型进行评估，计算模型在命名实体识别任务上的性能指标，如准确率、召回率、F1值等，分析模型在不同类型命名实体上的表现，以及对长短句子、不同上下文的适应能力。

# **实验数据**

## 数据获取

训练数据来自人民日报NER数据集，链接如下：

https://github.com/InsaneLife/ChineseNLPCorpus/tree/master/NER/renMinRiBao

## 数据统计

数据集由人民日报语料库1998版和2014版生成，包括人名(PER)、地名(LOC)和组织名(ORG)三类常见的实体类型，具体实体数没有明确说明。

# **方法/模型**

在本次研究中，我们基于Tensorflow深度学习框架，结合双向长短时记忆网络（BiLSTM）和条件随机场（CRF）算法来实现命名实体识别，对于给定文本中的命名实体标注人名、地名、组织名的标签。

双向长短时记忆网络是一种结合了双向循环神经网络和长短时记忆网络的神经网络模型。它能够更好地捕捉序列数据中的上下文信息，在自然语言处理、时间序列分析等领域都有着广泛的应用。

其中长短时记忆网络是一种特殊的循环神经网络，相比于普通的RNN，LSTM能够更好地解决长序列训练过程中的梯度消失和梯度爆炸问题。LSTM通过引入门控机制（遗忘门、输入门、输出门）来控制信息的流动，从而能够更好地捕捉长距离的依赖关系。

双向长短时记忆网络在LSTM的基础上增加了一个反向的LSTM结构，使得在每个时间步既可以利用当前时刻之前的信息（前向传播），也可以利用当前时刻之后的信息（反向传播）。这样，BiLSTM能够更全面地捕捉序列数据中的上下文信息，从而更好地解决一些序列建模的问题。

BiLSTM由前向LSTM和后向LSTM组成。在每个时间步，前向LSTM接受当前时刻的输入和前一时刻的隐藏状态，计算当前时刻的前向隐藏状态；后向LSTM接受当前时刻的输入和后一时刻的隐藏状态，计算当前时刻的后向隐藏状态。最终的输出可以由前向和后向隐藏状态的组合得到。在命名实体识别中，BiLSTM能够更好地理解和处理文本数据中的语境和上下文信息，从而提高模型的性能。

而条件随机场是一种用于序列标注问题的概率模型，它能够对给定的序列进行标注，并考虑到标注之间的依赖关系。CRF在自然语言处理、图像处理、生物信息学等领域都有广泛的应用。它将标注序列看作是一个随机变量序列，通过对标注序列的联合分布进行建模，来求解最优的标注序列。CRF可以捕捉到标注之间的依赖关系，从而能够更好地处理序列标注问题。CRF中的特征函数用于描述标注序列的特征，它将输入序列和标注序列映射到一个实数值。特征函数可以考虑到标注之间的依赖关系，从而提高模型的性能。CRF也是一种条件随机场，它的条件概率分布是给定输入序列条件下的标注序列的概率分布。CRF通过最大化条件概率分布来求解最优的标注序列。在命名实体识别中CRF能够考虑到标注之间的依赖关系，从而提高模型的性能。

综上所述，我们使用 BiLSTM 构建深度学习模型，同时考虑文本的前向和后向信息，以更全面地理解上下文，然后采用CRF对标注序列的联合分布进行建模，捕捉标注之间的依赖关系，从而更好地处理序列标注问题，提高NER模型的性能。

# **结果评测/展示**

## 执行概述

Train.py加载数据集、训练预测模型，Model文件夹中已包括训练好的模型，采用conda部署环境，安装requirement.txt中依赖包，执行predict.py即可。

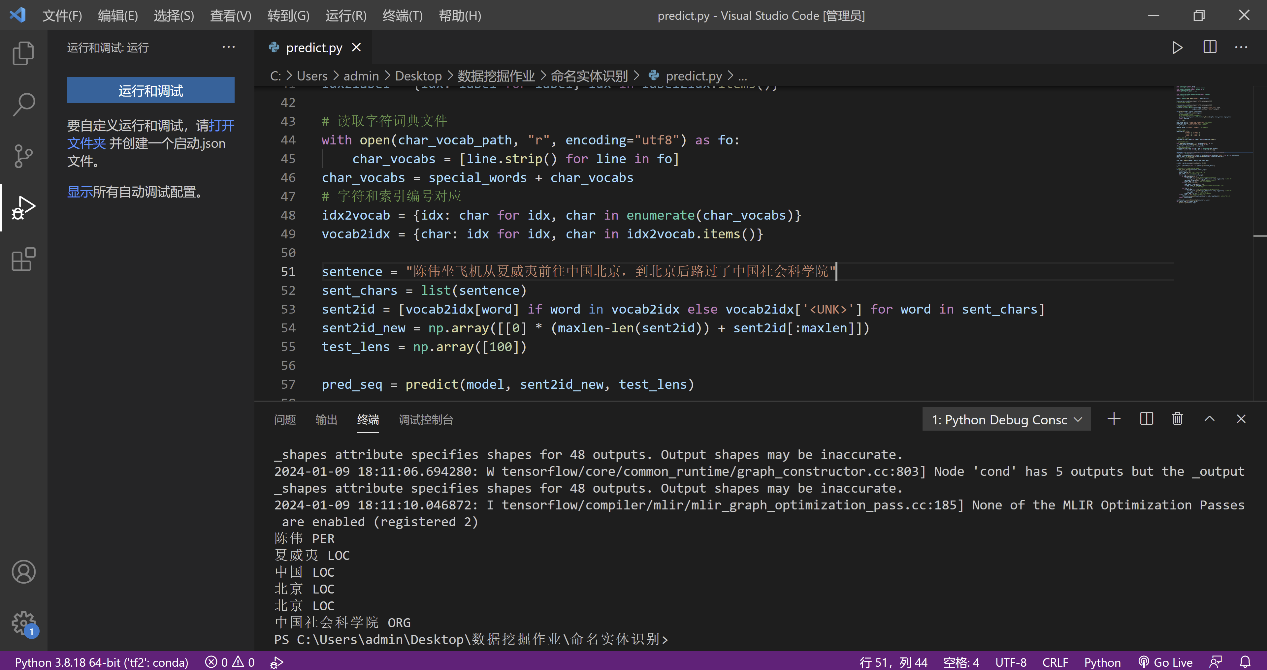
## 功能展示

下图展示了模型对sentence中句子“陈伟坐飞机从夏威夷前往中国北京，到北京后路过了中国社会科学院”进行命名实体识别的结果：

人名：陈伟

地名：夏威夷、中国、北京

组织名：中国社会科学院



# **总结**

## 局限性分析

命名实体的识别往往需要考虑上下文语境，尤其是对于实体与连接词、谓语动词一起出现的情况。比如“我和陈伟一起去故宫玩”，“陈伟”和“一起”会被模型识别为一个人名“陈伟一”，当前模型在对语境的理解上仍存在一定的局限性，在处理长距离依赖关系时表现尤为不佳。

在实际应用中，我们经常会遇到模型未曾见过的新实体，对于这些未知实体的识别和分类也是一个挑战。当前的模型在处理未知实体时表现较为不稳定，容易出现误识别或漏识别的情况，缺乏泛化能力。

进一步，命名实体的表达形式多样，包括简称、缩写、拼音、别名等，要求模型具备一定的多样性和多样化表达的能力。然而我们的模型在处理这种多样性表达的实体上存在诸多问题，还需要根据不同的领域和任务进行适应和迁移学习。

## 工作总结

在这次课程设计中，我们深入了解了双向长短时记忆网络和条件随机场的工作原理，利用TensorFlow深度学习框架，结合BiLSTM和CRF算法，实现了命名实体标签的初步预测。在这个过程中，我们认识到数据质量和标注的准确性对模型效果的影响至关重要，如果没有现成的数据集，获取大规模且高质量的标注数据将是一项消耗巨大的工作。另一方面，通过不断调试和修改，我们不仅提升了编写代码的能力，也对自然语言处理领域有了更深入的了解，受益匪浅。