表示学习

命名实体识别（NER）

背景知识

命名实体识别（Named Entity Recognition，简称NER）是自然语言处理（NLP）中的一项重要任务，旨在从文本中识别并分类具有特定意义的实体，如人名、地名、组织机构、日期、时间、百分比、货币等。NER的目标是识别文本中与特定类别相关的命名实体，并为这些实体分配预定义的类别标签。

NER系统通常采用监督学习方法，使用带有标签的训练数据进行模型训练。训练数据包含了文本句子和对应的实体标签，模型通过学习这些示例来预测新的文本中的实体和它们的类别。

基于ltp的命名实体识别

本次实验我们围绕命名实体识别任务展开，首先我们将先基于Python环境测试命名实体识别任务。
本次实验，我们首先使用由哈工大自主开发的中文语言分析工具LTP。LTP（Language Technology Platform）是由哈工大社会计算与信息检索研究中心（HIT-SCIR）开发的一套中文自然语言处理工具集。它提供了多个模块，涵盖了中文文本处理的各个方面，包括分词、词性标注、命名实体识别、依存句法分析等。

快速安装

```
pip install ltp
```

任务示例

运行示例代码，机器将下载对应已经训练好的ltp模型并进行加载。
同时，针对目标句子，模型将依次执行分词和命名实体识别两项任务，最终输出结果。

```
from ltp import LTP

ltp = LTP() #加载模型

sentence = "小明同学于今年暑假游览武汉。"
```

```
#两种任务：先分词、再进行命名实体识别
result = ltp.pipeline([sentence], tasks = ["cws","ner"])
print(result.ner)
```

运行得到结果如下：

```
[[('Nh', '小明'), ('Ns', '武汉')]]
```

ltp模型在训练时，其标注集如下所示：

命名实体识别标注集

LTP中的NE 模块识别三种NE，分别如下：

标记	含义
Nh	人名
Ni	机构名
Ns	地名

因此，在上述代码中对于例句进行命名识别时，其获得的输出为：('Nh','小明'),('Ns','武汉')

> 无法成功下载模型的同学，可以本地加载提供的模型：ltp = LTP("./ltp_small")

实验1

- 尝试更改例句，体验LTP工具对于不同实体的命名实体识别效果

基于bert的命名实体识别训练及测试实验

BERT是Transformer架构的一种变体，其创新在于引入了双向上下文理解，打破了传统自回归模型只能顺序处理信息的局限。通过Masked Language Modeling和Next Sentence Prediction两项任务，在无监督的数据上进行预训练，BERT学会了丰富的上下文语义表示。
在实际运用中，BERT模型被进一步微调以适应NER任务。通常，这涉及到在每个位置添加一个分类头，对每个输入的单词片段预测其所属的实体类型。模型利用标注数据进行训练，并通过优化损失函数最小化预测错误。

本次实验将基于哈工大讯飞联合实验室发布的基于全词Mask的中文预训练模型BERT-wwm（hfl/chinese-bert-wwm-ext），利用DUIE数据集进行微调，获得可以实现命名实体识别的预训练模型。（https://huggingface.co/hfl/chinese-bert-wwm-ext）

实验环境

包依赖

```
scikit-learn==1.1.3
scipy==1.10.1
seqeval==1.2.2
transformers==4.27.4
pytorch-crf==0.7.2
```

模型下载

下载预训练模型相关权重
参考网址：
https://huggingface.co/hfl/chinese-bert-wwm-ext
https://hf-mirror.com/hfl/chinese-bert-wwm-ext

数据处理

在Ner 任务中，数据通常需要被处理成BIO格式。BIO格式（Beginning, Inside, Outside）是一种用于自然语言处理任务中的标注方案，特别是在命名实体识别（NER）中广泛使用。BIO格式通过在每个词之前加上标签来标识该词在实体中的位置。具体来说，BIO格式包括以下标签：

1. **B-（Beginning）**：表示该词是一个实体的开始。
2. **I-（Inside）**：表示该词是一个实体的内部部分（不是开头）。
3. **O-（Outside）**：表示该词不属于任何实体。

例如：假设有一个句子“John lives in New York”，其中“John”和“New York”是需要识别的命名实体。这个句子的BIO格式标注可能是

Word	BIO Tag
John	B-PER
lives	O
in	O
New	B-LOC
York	I-LOC

在这个例子中：

- “John”是一个人名，因此标注为B-PER（开始一个人名实体）。
- “New”是一个地名的开始部分，因此标注为B-LOC（开始一个地名实体）。
- “York”是一个地名的内部部分，因此标注为I-LOC（地名实体的内部部分）。
- “lives”和“in”不是任何实体的一部分，因此标注为O（Outside）。

实验2 数据处理

- 基于提供的data文件夹中下的 dev.json, test.json, train.json, duie_schema.json的数据文件完成后续实验。
- 补全 process.py 文件Todo部分，实现 dev.json, train.json文件向BIO文件的转换并保存。

模型训练

模型的训练和保存涉及四个文件：

1. model.py 用于定义模型相关结构和基于输入的生成、推理函数
2. main.py 模型的训练checkpoint保存代码
3. data_loader.py 用于数据集加载
4. config.py 模型训练相关配置文件

实验3 模型训练和保存

- 补全model.py、main.py、data_loader.py的Todo部分，实现模型的训练和权重保存，可以根据硬件配置调整训练设置和训练集、测试集大小，加速训练过程，其中config中的默认配置参数为：

```
max_seq_len=256
train_batch_size=12
dev_batch_size=12
save_step=500
epochs=1
```

训练成功样例

```
【train】1/3 1/2502 loss:273.1101379394531
【train】1/3 2/2502 loss:247.451171875
【train】1/3 3/2502 loss:259.4442443847656
【train】1/3 4/2502 loss:353.6705017089844
【train】1/3 5/2502 loss:238.2313995361328
【train】1/3 6/2502 loss:224.66835021972656
【train】1/3 7/2502 loss:230.4559783935547
【train】1/3 8/2502 loss:185.7163543701172
【train】1/3 9/2502 loss:122.71273040771484
```

实验4 NER模型测试

- 基于训练得到的权重编写模型测试代码，在测试数据集上完成测试，提供完整版模型预测代码并提供以下例句的测试结果截图：

```
《民航客运服务会话》是1995年中国民航出版社出版的图书，作者是周石田
再有之后的《半生缘》，蒋勤勤饰演的顾曼璐完全把林心如的曼桢衬得像是涉世未深的小姑娘，毫无半点风情
裴友生，男，汉族，湖北蕲春人，1957年12月出生，大专学历
吴君如演的周吉是电影《花田喜事》，在周吉大婚之夜，其夫林嘉声逃走失踪，后来其夫新科状元高中回来，周吉急往城楼相识，但林嘉声却言夫妻情断，覆水难收
```

基于大模型的命名实体识别

大语言模型在训练时通过大量的文本数据学习了丰富的语言结构和上下文信息。这使得模型能够更好地理解命名实体在文本中的上下文，提高了识别的准确性。即使模型在训练过程中没有见过某个命名实体，它也可以通过上下文推断该实体的类别。这意味着模型可以处理新的、未知的实体，而无需重新训练。经过指令微调的大模型，具有较强的指令跟随能力，用户可以通过输入指令来促使大模型执行各种复杂任务。

实验5 基于大模型的命名实体识别

- 尝试编写指令，使用国产大语言模型对于实验3中的测试例句进行命名实体识别。要求模型按照json格式输出对应结果。
  参考大模型：

文心一言：https://yiyan.baidu.com/
通义千问：https://qianwen.aliyun.com/