实验2

命名实体识别

前言

命名实体识别是NLP领域的一项基础且重要的任务，它旨在将一串文本中的实体识别出来，并标注出它所指代的类型，比如人名、地名等等。具体地，根据MUC会议规定，命名实体识别任务包括三个子任务：

- 实体名：人名、地名、机构名等
- 时间表达式：日期、时间、持续时间等
- 数字表达式：百分比、度量衡、钱、基数等

来看这句话，百度于2021年3月23日正式回香港上市，这句话百度是个机构名，香港是个地名，2021年3月23日是个日期，命名实体识别任务能够通过建模的方式来帮助自动地发现这些实体。

- **实验数据**

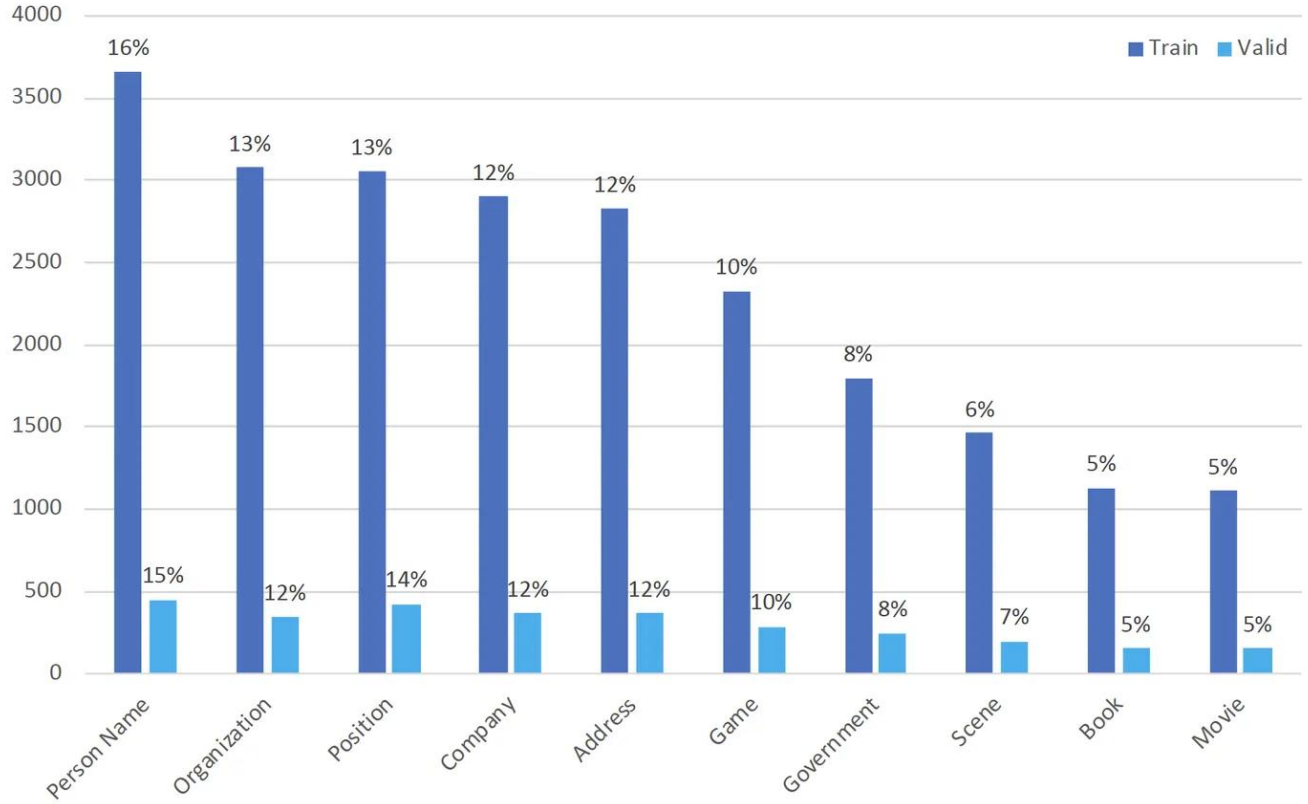
原始数据存储在位于具体模型的/data/clue/路径下，train.json和test.json文件中。文件中的每一行是一条单独的数据，一条数据包括一个原始句子以及其上的标签，具体形式如下：

```
{
        "text": "浙商银行企业信贷部叶老桂博士则从另一个角度对五道门槛进行了解读。叶老桂认为，对目前国内商业银行而言，",
        "label": {
                "name": {
                        "叶老桂": [
                                [9, 11],
                                [32, 34]
                        ]
                },
                "company": {
                        "浙商银行": [
                                [0, 3]
                        ]
                }
        }
}
```

• 实验数据

实验数据来自CLUENER2020，这是一个中文细粒度命名实体识别数据集，是基于开源文本分类数据集THUCNEWS，选出部分数据进行细粒度标注得到的。该数据集的训练集、验证集和测试集的大小分别为10748，1343，1345，平均句子长度37.4字，最长50字。

CLUENER2020共有10个不同的类别，包括：组织(organization)、人名(name)、地址(address)、公司(company)、政府(government)、书籍(book)、游戏(game)、电影(movie)、职位(position)和景点(　　)、



图为10种标签在训练集和原始验证集上的分布情况，可以看到标签样本不均衡的问题比较明显，人名实体的标签是电影类标签的近三倍多，这也给模型训练带来了挑战。

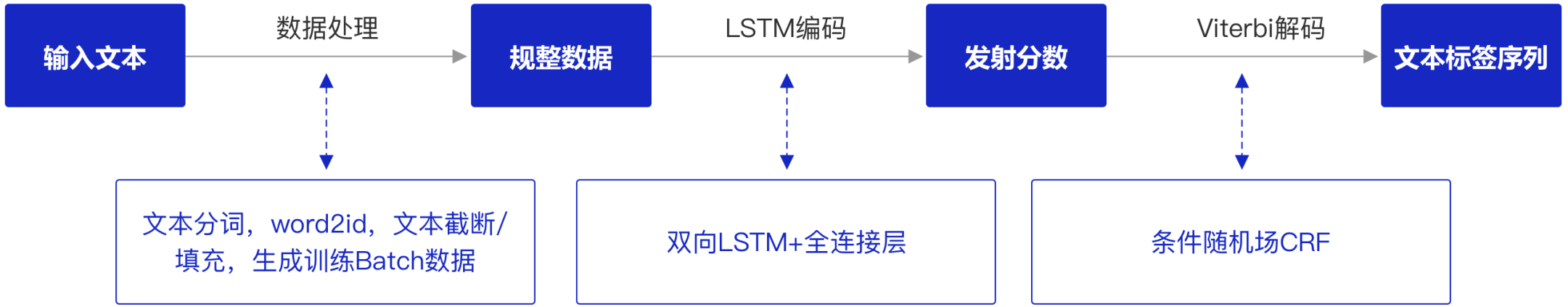
- # 实验设计

命名实体识别任务会被建模成序列标注任务，也就是说，模型的输入是待识别的一串文本序列，模型的输出就是该文本序列对应的标签序列，是一种序列到序列的任务。

姚	明	担	任	中	国	篮	协	主	席
B-person	I-person	O	O	B-organization	I-organization	I-organization	I-organization	O	O

这句话中的每个字分别对应着一个标签， 模型的输入就是上边的文本，模型的输出就是下面的标签序列，通过这样的标签序列就能识别出上边文本中的实体。
具体来说，上边这串文本中，"姚明"对应着person实体，其中"姚"字是"person"实体的起始字，所以设置标签为"B-person"，其中标签前边的B代表Begin这个单词；"明"字是"person"实体的中间字，所以设置标签为"I-person"，其中标签前边的I代表Inside这个单词。 "中国篮协"对应这organization实体，相应标签"B-organization"和"I-organization"的解读和person实体是一致的。最后的标签"O"代表"Outside"，表示其他实体类型的标签。

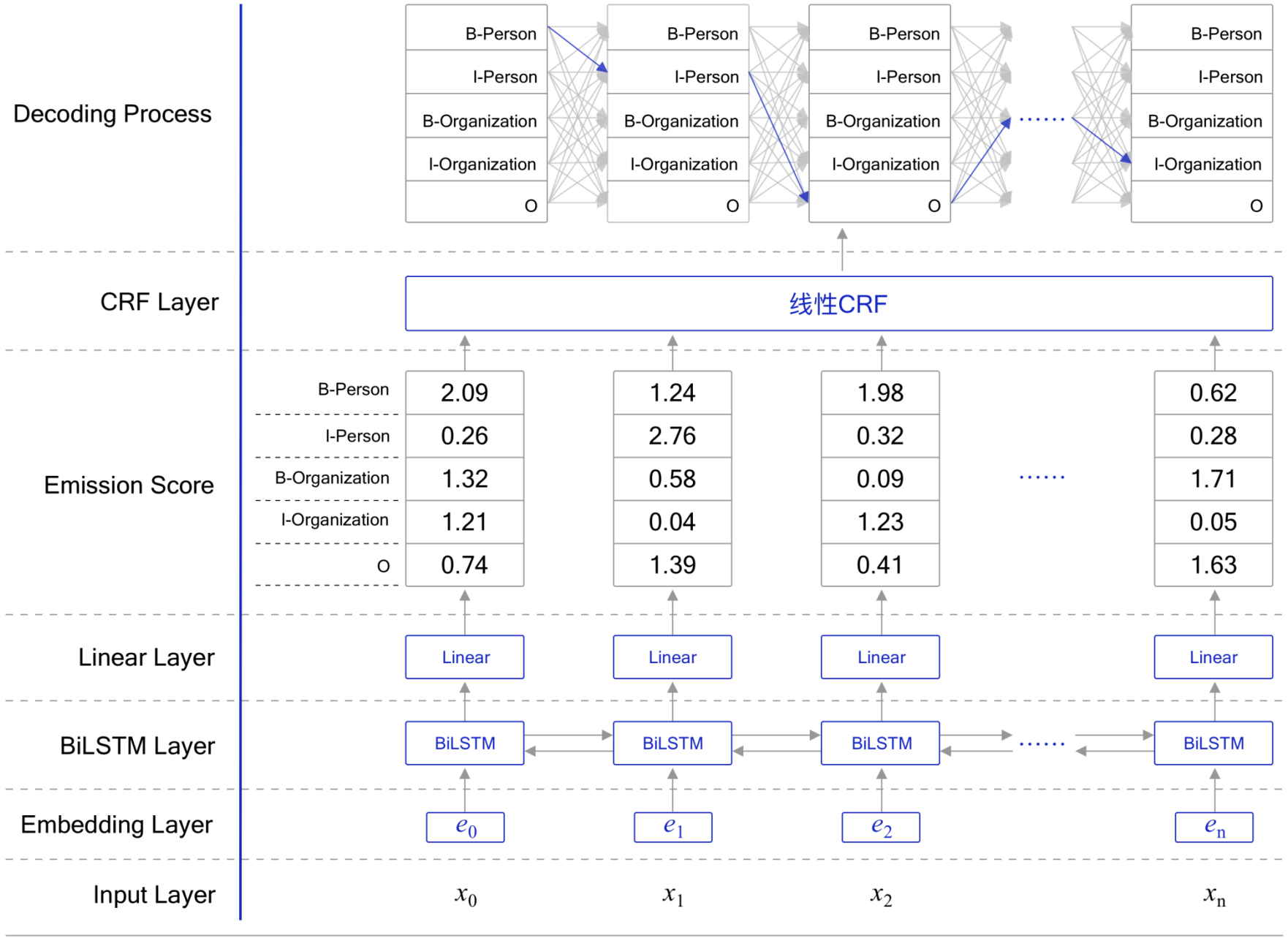
- 模型设计



利用BiLSTM+CRF的结构来实现命名实体识别任务
模型的输入是待识别的文本序列，模型的输出就是该文本对应tag序列。在建模过程中，对于输入的待识别文本，首先需要进行数据处理生成规整的文本序列数据，包括语句分词、将词转换为id，过长文本截断、过短文本填充等等操作；然后，将文本序列传入到双向的LSTM模型中，这样在LSTM每个时间步骤都能获得一个与输入对应的向量，然后将这些向量传入全连接层，将会得到一个被称为"发射分数"的向量。最后，将这个发射分数传给条件随机场CRF，CRF会根据这个"发射分数"进行解码，得到原始输入对应的标签序列。

- 实验理论解读

假设当前有一串长度为n的文本x=[$x1, x2, x3, ..., xn$]，其中xi代表该文本序列中的第i个字。另外，假设实体词类型一共有2个，分别是person和organization实体，这样对应标签的数量一共有5个，分别是B-person, I-person, B-organzaiton, I-organization, O。



- 实验实现

1. 环境：Python 3.6+ and PyTorch 1.5.1.
   pip install -r requirements.txt
2. 需要的预训练模型：pytorch_model.bin vocab.txt
放置在./pretrained_bert_models对应的预训练模型文件夹下，其中bert-base-chinese模型：pytorch版本的模型放在下载链接里
3. 运行：run.py，模型运行结束后，最优模型和训练log保存在./experiments/clue/路径下。在测试集中的bad case保存在./case/bad_case.txt中。
注意：当前模型的train.log已保存在./experiments/clue/路径下，如要重新运行模型，请先将train.log移出，以免覆盖。

- **资源下载**

下载代码，成功复现，可以解释代码（6-7分）

- 完善代码

补充其他命名实体识别模型（每个模型1分）

CLUE命名实体任务排行榜 项目地址 | 论文 | 数据集与提交样例

排行	模型	研究机构	测评时间	Score	认证	CLUENER
1	IE-Model	北京百分点科技集团股份有限公司	23-05-12	83.599	待认证	83.599
2	CLUENER	test	23-04-20	83.400	待认证	83.400
3	JoveBM	GTCOM-NLP	22-11-10	83.368	待认证	83.368
4	OBERT	OPPO小布助手	22-07-12	83.351	待认证	83.351
5	bert_span	lier	22-07-11	83.336	待认证	83.336
6	DML_nezha_large	sensetime-scg-stc2-nlp-group	22-04-27	83.308	待认证	83.308
7	roformer-large	sensetime-nlp	22-08-28	83.276	待认证	83.276
8	kg_le_roberta_crf_ensemble	ant_group_postlending_ai	22-04-12	83.254	待认证	83.254
9	MacBERT	CMB AI Lab	22-04-10	83.247	待认证	83.247
10	K-BERT+Roberta+Nezha+Mengzi	Only	22-11-10	83.221	待认证	83.221

BERT-CRF
BERT-LSTM-CRF
BERT-Softmax

https://github.com/CLUEbenchmark/CLUENER2020

- 完善代码

结果分析：

模型	BiLSTM+CRF	Roberta+Softmax	Roberta+CRF	Roberta+LSTM+CRF
address				
book				
company				
game				
government				
movie				
name				
organization				
position				
scene				
overall				

谢谢

FOR YOUR LISTENING

陈雅妮.