中文命名实体识别——自然语言处理实验报告

DZ1833003 曹雨露

实验目标

中文命名实体识别——从已经分好词的中文文本中识别出人名（PERSON）、地点（LOCATION）、时间（TIME）及机构名（ORGANIZATION）。

实验方法

主要方法：双向LSTM+CRF模型实现中文命名实体标注

- 预先数据处理：
  - 词向量：使用word2vec训练得到词向量
  - 数据字典：由训练数据train.txt得到word2id,id2word,tag2id,id2tag，x_train,y_train储存进data_train.pkl
  - 验证集和测验集：由数据字典word2id，tag2id，生成一定格式的data_dev.pkl,data_test.content.pkl
- 建立LSTM+CRF模型
  - 调用tnsorflow的双向LSTM神经网络模块，进行学习，双向LSTM同时考虑了过去的特征（通过前向过程提取）和未来的特征（通过后向过程提取），将双向LSTM神经网络的输出拼接，该输出作为后接的CRF的输入，计算获取全局最优的输出序列
  - 超参数：lr（学习率），dropout_keep（下降率）
- 输入数据
  一轮训练包含多组batch_size的数据作为输入，一次输入是从训练数据中取出长度为batch_size的数据，数据每一行padding为最大长度，feed进模型中进行学习，需要feed进的数据包含训练数据(seqs)、标签数据(labels)，句子长度的列表(seqs_list)
- 执行训练
  设置epoch，训练多轮，每一轮保存模型和模型数据，并在验证集上执行一次验证，得到每一次的f_measure

实验参数

- epochs = 10
- batch_size = 32
- lr = 0.001
- embedding_dim = 100 #词向量为维度
- dropout_keep = 0.5 #训练时为0.5

运行环境和运行方式

运行环境

Python3.6.5
tensorflow (1.12.0)

在不同模式下的使用方法

- 训练

```
python3 train.py pretrained
```

在训练过程中，每训练一轮会执行一次验证

- 验证

```
python3 train.py dev
```

- 测试

```
python3 train outputpath
```

outputpath为输出数据路径，如data/DZ1833003.txt

实验结果

训练得到的模型在验证集上的性能f1-measure-overall能够达到0.90 ~ 0.91，在测验集上则为0.899662

实验总结

双向LSTM+CRF模型在命名实体标注上已经取得了较好的结果，参数会对模型有一定影响，在本次实验中，我学习了如何使用tensorflow，对自然语言处理的应用有了更深的认识，同时也遇到了很多问题，在参考了github上的项目之后发现了许多问题，如模型中的参数不应该是固定维度，而是应该是动态的等，最终解决了这些问题，完成了实验任务。

参考内容

https://blog.csdn.net/jmh1996/article/details/83476061
https://github.com/buppt/ChineseNER