训练深度学习命名实体抽取模型

目前实现了端对端的深度学习命名实体抽取模型，包括CNN+ATT+LSTM+CRF。不需要人工特征抽取，可以达到总体准确率89%左右。

代码地址

https://github.com/yefengwang/medical_ner

下载代码

$git clone https://github.com/yefengwang/medical_ner

生成训练数据集

ehost2bio.py 文件可以将ehost项目文件转化为bio格式，并生成train, validation和test数据。

$python ehost2bio.py -c ehost_workspace/20170818 work

假定ehost_workspace/20170818目录内包含所有的标注

生成一个work文件夹，内产生3个文件，分别为train.txt, valid.txt, test.txt。

train.txt里面有70%的数据，作为训练集，valid.txt里面有15%的数据为调试集，test.txt内有15%数据为测试集。

生成训字典

字典由训练集里面的字和词组成，用来做特征索引

$python NERModel.py -b work

将会在work目录下生成如下文件：

label_vocab.txt 标签字典，BIO标签
char_vocab.txt 字字典，包含训练集内的每一个字
word_vocab.txt 词字典 ，包含训练集内的每一个词
word.npz 词embeddings，根据embeddings/word.txt生成
char.npz 字embeddings，根据embeddings/char.txt生成

训练模型

$python NERModel.py work

将会进行模型训练，结果写入work/model目录

模型参数

这些参数定义在NERModel.py开头，可以依照需要进行调整

```
# Hyper parameters
word_embedding_size = 100 # word embedding size
char_embedding_size = 100 # char embedding size
kernels = [2, 3] # CNN 过滤器，为字的长度，2为2个字，3为三个字
cnn_hidden_size = 200 # char CNN 层的最终大小
lstm_hidden_size = 300 # LSTM 层的大小

converge_check = 3 #如果3次迭代模型没有提高则停止训练，为了防止过度拟合
use_chars = True #使用char-cnn模型
use_crf = True #使用CRF解码
use_char_attention = True #使用attention模型来连接词特征与字特征
clip = 5 #防止gradient vanish
batch_size = 20
num_epochs = 35 #迭代次数
dropout = 0.5
learning_rate = 0.001
learning_rate_decay = 0.9
```

预测

$python prediction.py work ehost_workspace/20170818/corpus/0001.txt

对0001.txt文件进行预测，输出bio文件格式。

work目录为模型与字典所在目录。

交叉验证

目的在于在全部数据集上测试模型准确性

类似于之前的步骤，只是执行每一步的时候用不同的参数。

生成 10-fold cross validation数据
$python ehost2bio.py -f 10 ehost_workspace/20170818 work

训练10 fold 交叉验证模型
$python NERModel.py -f 10 work

将结果复制到 cnn-lstm-crf目录中
$python fold_concat.py work cnn-lstm-crf

显示交叉验证结果
$python fold_summary.py cnn-lstm-crf

结果

结果为cnn+attention+lstm+crf 模型，目前没有使用word embeddings

	Correct	Answer	Actual	Precision	Recall	F-score
测量数值	478	525	522	91.05	91.57	91.31
检验	193	269	263	71.75	73.38	72.56
检查	1354	1483	1562	91.30	86.68	88.93
药物	71	126	142	56.35	50.00	52.99
时间词	161	195	218	82.56	73.85	77.97
治疗	334	422	412	79.15	81.07	80.10
疾病	924	1123	1096	82.28	84.31	83.28
身体部位	3167	3467	3381	91.35	93.67	92.49
医学发现	5998	6580	6645	91.16	90.26	90.71
Overall	12680	14190	14241	89.36	89.04	89.20