

作业二报告

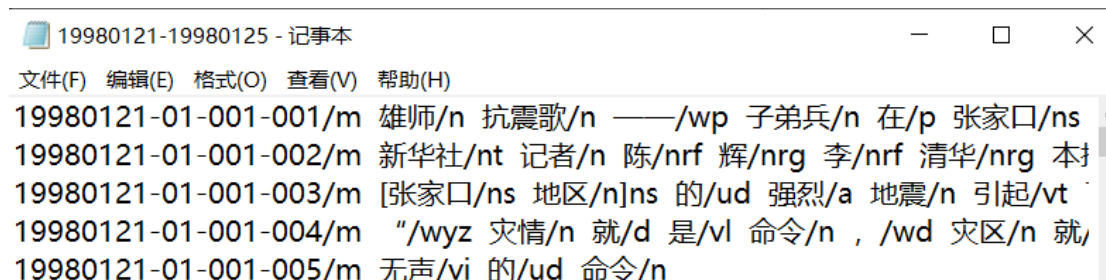
1 概述

本次作业对 1998 年 1 月人民日报做了命名实体识别，识别对象为机构名，使用了词嵌入表示并采用 softmax 实现多分类识别 BIO 标签。在实验中词嵌入为已训练完成的网络资源，其维度为 50，梯度下降采用小批量梯度下降，相较于作业 1，F1-measure 略有下降。

2.实验流程

2.1 分割原始数据

将原文件重命名为“renming.txt”，根据每行开头日期将其划分为三个 txt 文件，“19980101-19980120.txt”、“19980121-19980125.txt”、“19980126-19980131.txt”，分别作为训练集、验证集、测试集的初始文本文件。该步骤运行 data 文件夹下 split_collection.py 即可完成，运行结果符合预期，以下是“19980121-19980125.txt”文件部分截图：



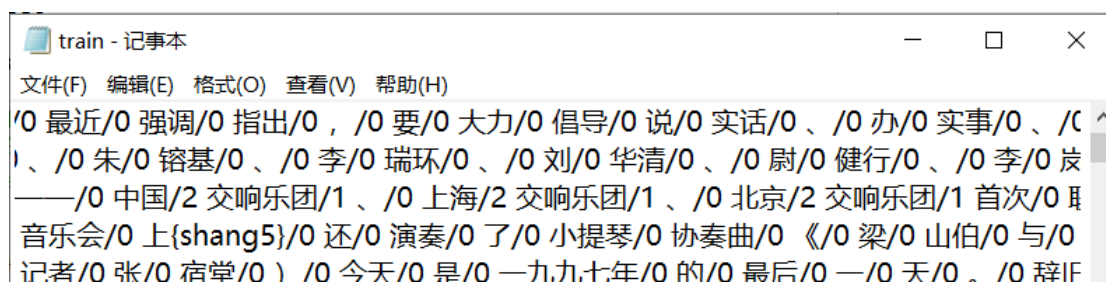
```
19980121-01-001-001/m 雄师/n 抗震歌/n ——/wp 子弟兵/n 在/p 张家口/ns
19980121-01-001-002/m 新华社/nt 记者/n 陈/nrf 辉/nrg 李/nrf 清华/nrg 本
19980121-01-001-003/m [张家口/ns 地区/n]ns 的/ud 强烈/a 地震/n 引起/vt
19980121-01-001-004/m "/wyz 灾情/n 就/d 是/vl 命令/n , /wd 灾区/n 就,
19980121-01-001-005/m 无声/vi 的/ud 命令/n
```

2.2 预处理训练集、验证集、测试集

根据获得的常用词典转化训练集、验证集、测试集，减少此后训练时的内存消耗。

分别转化三个集合，将每个词转化为 word1/word2 的形式。读取集合文件，对于每一个词判断其词性，若为机构名命名实体的首个词则标记 word2 为 2，并将此后该命名实体的其他词 word2 标记为 1，其余情况均标记为 0，具体操作方法是创建临时列表，非[]内词直接判断，[]内词需记录[]内所有词后一并修改，并根据[]词性判断词性。

本步骤运行 predeal.py 即可完成，以下为“train.txt”文件部分截图：



```
/0 最近/0 强调/0 指出/0 , /0 要/0 大力/0 倡导/0 说/0 实话/0 、 /0 办/0 实事/0 、 /0
/0 朱/0 镕基/0 、 /0 李/0 瑞环/0 、 /0 刘/0 华清/0 、 /0 尉/0 健行/0 、 /0 李/0 岚
——/0 中国/2 交响乐团/1 、 /0 上海/2 交响乐团/1 、 /0 北京/2 交响乐团/1 首次/0 且
音乐会/0 上{shang5}/0 还/0 演奏/0 了/0 小提琴/0 协奏曲/0 《/0 梁/0 山伯/0 与/0
记者/0 张/0 宿堂/0 \ /0 今天/0 是/0 一九九七年/0 的/0 最后/0 一/0 天/0 。 /0 辞IF
```

2.3 训练模型算 F1-measure

根据预处理过后的训练集，建立 softmax 模型进行梯度下降，本次作业采取梯度下降的方式，每次迭代使用 batch 个样本对 theta 进行更新（本次作业 batch 设为 128），为确保效率，每对训练集所有样本遍历一次计算一次验证集上的 F1-measure。

首先需要根据已有的词向量文件建立字典方便数据处理。读取预处理数据通过自定义类（需继承 torch 库中 Dataset）来完成，该类需包含二维列表，记录对应集合中每个样本对应的数据（由该样本前中后三个词的词向量连接成的[3*size,1]tensor 张量），Softmax 模型需设置正向传播函数，为计算方便所以为返回值添加 log 操作，配合 torch 库中 NLLLoss 损失，实现对模型的训练。F1-measure 的计算通过统计真实值和预测值中命名实体来进行，即将某一命名实体的起始序号和终止序号作为元组添加至对应的集合中，真实集和预测集的交集的大小即为 TP，即可获得查准率和查全率，进一步计算 F1-measure。

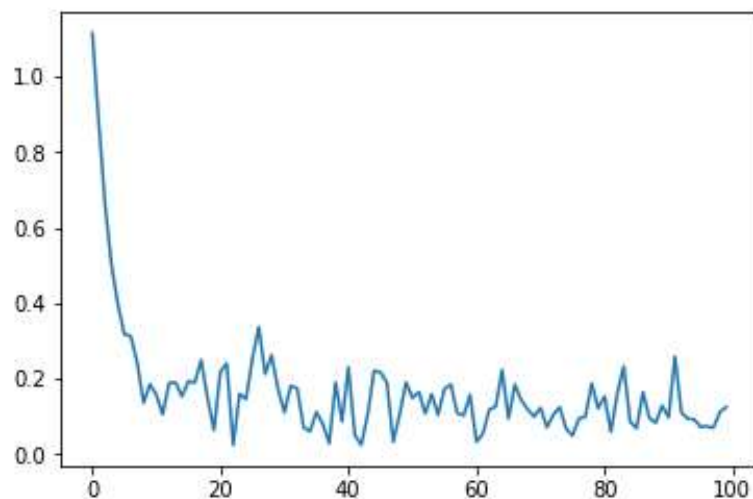
实例化 softmax 模型，设定损失函数，定义优化器，并使用 torch 库中 Dataloader 设定迭代方式 (batch 大小、随机化处理)，设定 epochs 参数，一个 epoch 表明训练集中每个样本都参与训练一次，每经过一次 epoch 计算一次 F1-measure，每次 epoch 包含多次迭代操作，每次迭代取 batch 个样本对模型进行更新，若取到训练集最后不足 batch 个样本则取小于 batch 个样本进行更新。根据样本和模型预测结果，并用损失函数计算 loss，清零梯度后对 loss 进行反向传播，并更新参数，即完成一次 batch。

每次 epoch 后计算验证集 F1-measure 值。读入验证集文件，并根据模型进行预测，并形成验证集的命名实体预测集，和原有的命名实体真实集进行交操作即可获得 TP 值，查准率为 $TP/(预测集大小)$ ，查全率为 $TP/(真实集大小)$ ，进而计算 F1-measure。

运行 main.py,修改参数即可进行训练并得到，并生成 loss 下降图。

3.结果及分析

本次作业采取维数为 50 的词向量，并使用 softmax 模型对机构名命名实体做 BIO 多分类处理，可以在较短的时间内获得结果，但相较于第一次作业测试集 F1-measure 达到 0.6165 的结果有明显下降，且在多次训练中有较大的起伏，验证集 F1-measure 在 0.33 至 0.38 间波动，对应测试集 F1-measure 为 0.3404，以下为 loss 值在 100 次更新的结果图：



观察可知 loss 很快下降至约 0.2，并在[0.0,0.4]区间内波动。

4.总结

原本是打算根据第一次作业代码改写的，但第一次作业仅仅写了手动求导的版本，没有深入了解 pytorch 的各项功能，只能重构代码，也算是让原来的代码规范了不少，也方便此后根据需求修改代码。

此次作业得出的结果相较第一次并无提升，甚至有明显的下降，虽然一方面是因为提高了识别命名实体的标准，但另一方面也可能是词向量和 softmax 模型本身的问题，词向量特征本身是非线性的，但模型是线性模型，使用线性函数拟合非线性函数的结果可能并不会十分良好。